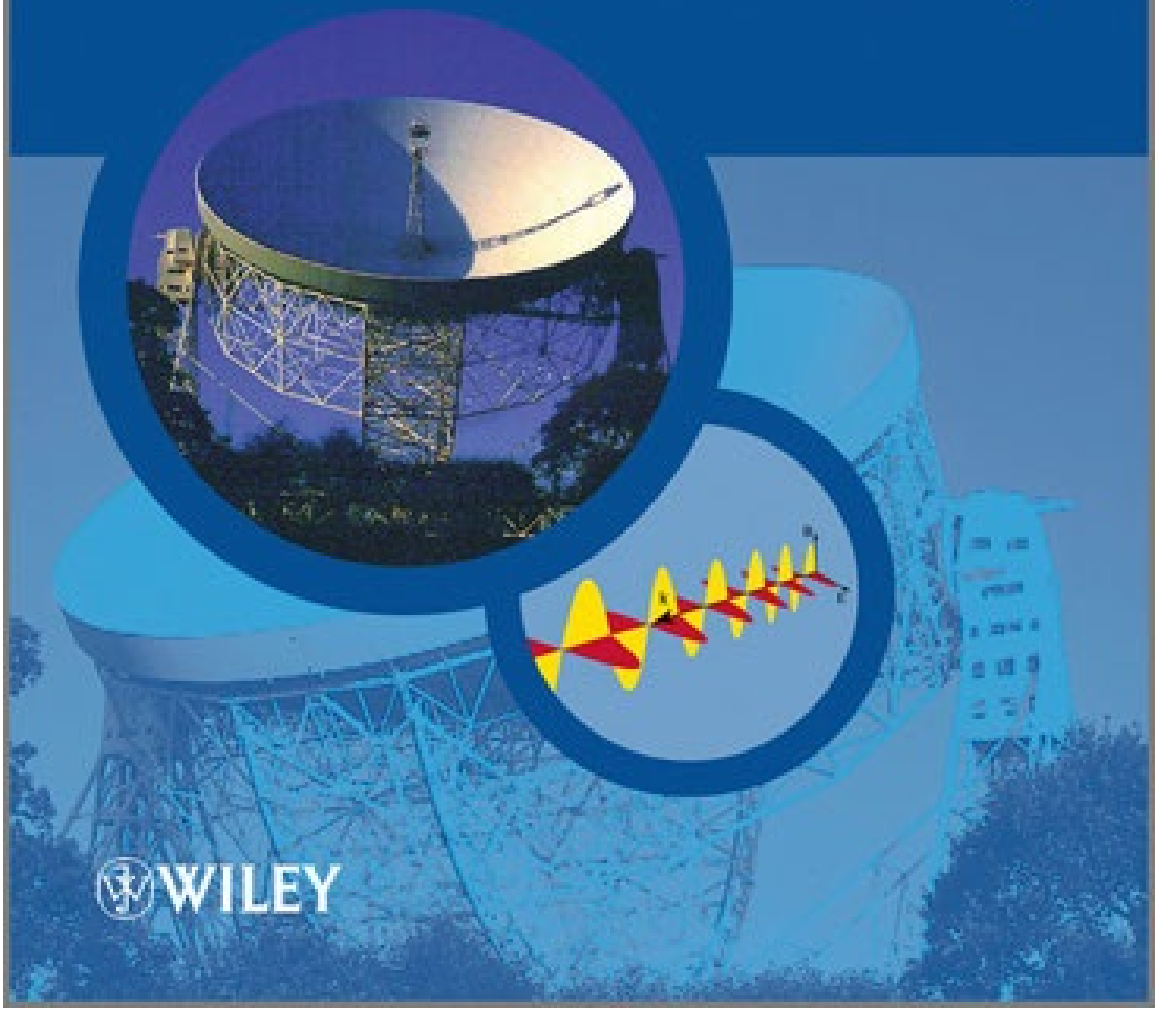
Electro-. magnetism

SECOND EDITION

I. S. Grant & W. R. Phillips

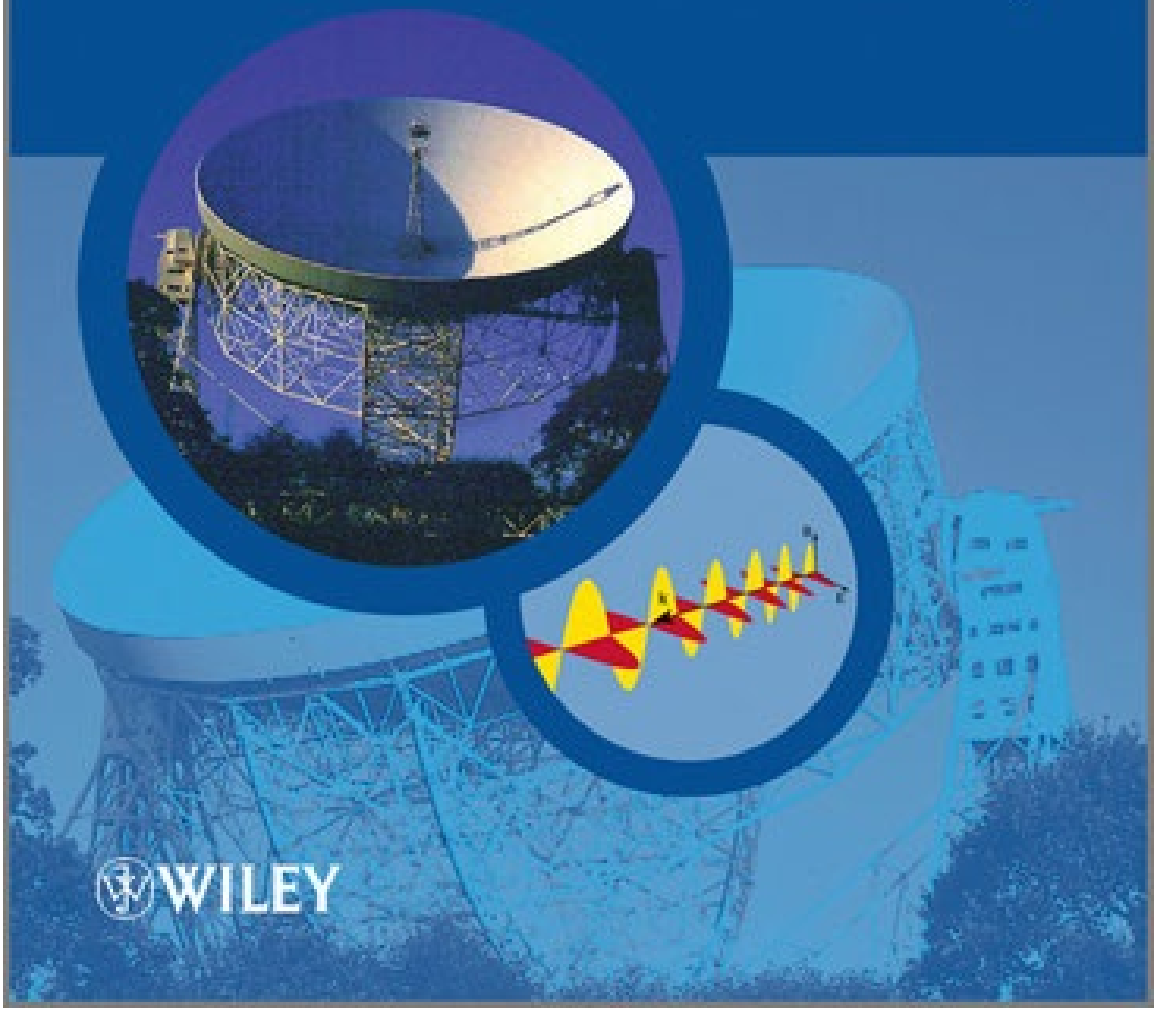


WILEY

Electro-magnetism

SECOND EDITION

I. S. Grant & W. R. Phillips



WILEY

Contents

CHAPTER 1 Force and energy in electrostatics

1.1 ELECTRIC CHARGE

1.2 THE ELECTRIC FIELD

1.3 ELECTRIC FIELDS IN MATTER

1.4 GAUSS' LAW

1.5 ELECTROSTATIC ENERGY

PROBLEMS 1

CHAPTER 2 Dielectrics

2.1 POLARIZATION

2.2 RELATIVE PERMITTIVITY AND ELECTRIC SUSCEPTIBILITY

2.3 MACROSCOPIC FIELDS IN DIELECTRICS

2.4 ENERGY IN THE PRESENCE OF DIELECTRICS

PROBLEMS 2

CHAPTER 3 Electrostatic field calculations

3.1 POISSON'S EQUATION AND LAPLACE'S EQUATION

3.2 BOUNDARIES BETWEEN DIFFERENT REGIONS

3.3 BOUNDARY CONDITIONS AND FIELD PATTERNS

3.4 ELECTROSTATIC LENSES

3.5 NUMERICAL SOLUTIONS OF POISSON'S EQUATION

3.6 SUMMARY OF ELECTROSTATICS

PROBLEMS 3

CHAPTER 4 Steady currents and magnetic fields

4.1 ELECTROMOTIVE FORCE AND CONDUCTION

4.2 THE MAGNETIC FIELD

4.3 THE MAGNETIC DIPOLE

4.4 AMPERE'S LAW

4.5 THE DIFFERENTIAL FORM OF AMPERE'S LAW

4.6 FORCES AND TORQUES ON COILS

4.7 THE MOTION OF CHARGED PARTICLES IN ELECTRIC AND MAGNETIC FIELDS

PROBLEMS 4

CHAPTER 5 Magnetic materials

5.1 MAGNETIZATION

5.2 THE MACROSCOPIC MAGNETIC FIELD INSIDE MEDIA

5.3 THE FIELD VECTOR H

5.4 MAGNETS

5.5 SUMMARY OF MAGNETOSTATICS

PROBLEMS 5

CHAPTER 6 Electromagnetic induction and magnetic energy

6.1 ELECTROMAGNETIC INDUCTION

6.2 SELF-INDUCTANCE AND MUTUAL INDUCTANCE

6.3 ENERGY AND FORCES IN MAGNETIC FIELDS

6.4 THE MEASUREMENT OF MAGNETIC FIELDS AND SUSCEPTIBILITIES

PROBLEMS 6

CHAPTER 7 Alternating currents and transients

7.1 ALTERNATING CURRENT GENERATORS

7.2 AMPLITUDE, PHASE AND PERIOD

7.3 RESISTANCE, CAPACITANCE AND INDUCTANCE IN A.C. CIRCUITS

7.4 THE PHASOR DIAGRAM AND COMPLEX IMPEDANCE

7.5 POWER IN A.C. CIRCUITS

7.6 RESONANCE

7.7 TRANSIENTS

PROBLEMS 7

CHAPTER 8 Linear circuits

8.1 NETWORKS

8.2 AUDIO-FREQUENCY BRIDGES

8.3 IMPEDANCE AND ADMITTANCE

8.4 FILTERS

8.5 TRANSFORMERS

PROBLEMS 8

CHAPTER 9 Transmission lines

9.1 PROPAGATION OF SIGNALS IN A LOSSLESS TRANSMISSION LINE

9.2 PRACTICAL TYPES OF TRANSMISSION LINE

9.3 REFLECTIONS

9.4 THE INPUT IMPEDANCE OF A MISMATCHED LINE

9.5 LOSSY LINES

PROBLEMS 9

CHAPTER 10 Maxwell's equations

10.1 THE EQUATION OF CONTINUITY

10.2 DISPLACEMENT CURRENT

10.3 MAXWELL'S EQUATIONS

10.4 ELECTROMAGNETIC RADIATION

10.5 THE MICROSCOPIC FIELD EQUATIONS

PROBLEMS 10

CHAPTER 11 Electromagnetic waves

11.1 ELECTROMAGNETIC WAVES IN FREE SPACE

11.2 PLANE WAVES AND POLARIZATION

11.3 DISPERSION

11.4 ENERGY IN ELECTROMAGNETIC WAVES

11.5 THE ABSORPTION OF PLANE WAVES IN CONDUCTORS AND THE SKIN EFFECT

11.6 THE REFLECTION AND TRANSMISSION OF ELECTROMAGNETIC WAVES

11.7 ELECTROMAGNETIC WAVES AND PHOTONS

PROBLEMS 11

CHAPTER 12 Waveguides

12.1 THE PROPAGATION OF WAVES BETWEEN CONDUCTING PLATES

12.2 RECTANGULAR WAVEGUIDES

12.3 CAVITIES

PROBLEMS 12

CHAPTER 13 The generation of electromagnetic waves

13.1 THE RETARDED POTENTIALS

13.2 THE HERTZIAN DIPOLE

13.3 ANTENNAS

PROBLEMS 13

CHAPTER 14 Electromagnetism and special relativity

14.1 INTRODUCTORY REMARKS

14.2 THE LORENTZ TRANSFORMATION

14.3 CHARGES AND FIELDS AS SEEN BY DIFFERENT OBSERVERS

14.4 FOUR-VECTORS

14.5 MAXWELL'S EQUATIONS IN FOUR-VECTOR FORM

14.6 TRANSFORMATION OF THE FIELDS

14.7 MAGNETISM AS A RELATIVISTIC PHENOMENON

14.8 RETARDED POTENTIALS FROM THE RELATIVISTIC STANDPOINT

PROBLEMS 14

APPENDIX A Units

A.1 ELECTRICAL UNITS AND STANDARDS

A.2 GAUSSIAN UNITS

A.3 CONVERSION BETWEEN SI AND GAUSSIAN UNITS

APPENDIX B Fields and differential operators

B.1 THE OPERATORS DIV, GRAD AND CURL

B.2 FORMULAE IN DIFFERENT COORDINATE SYSTEMS

B.3 IDENTITIES

APPENDIX C The derivation of the Biot-Savart law

Solutions to Problems

Further Reading

Index

The Manchester Physics Series

General Editors

D. J. SANDIFORD: F. MANDL: A. C. PHILLIPS

Department of Physics and Astronomy, Faculty of Science,

University of Manchester

Properties of Matter:	B. H. Flowers and E. Mendoza
Optics: Second Edition	F. G. Smith and J. H. Thomson
Statistical Physics: Second Edition	F. Mandl
Electromagnetism: Second Edition	I. S. Grant and W. R. Phillips
Statistics:	R. J. Barlow
Solid State Physics: Second Edition	J. R. Hook and H. E. Hall
Quantum Mechanics:	F. Mandl
Particle Physics: Second Edition	B. R. Martin and G. Shaw
The Physics of Stars: Second Edition	A. C. Phillips
Computing for Scientists:	R. J. Barlow and A. R. Barnett

ELECTROMAGNETISM

Second Edition

I. S. Grant
W. R. Phillips

Department of Physics
University of Manchester

John Wiley & Sons

CHICHESTER NEW YORK BRISBANE TORONTO SINGAPORE

Copyright © 1975, 1990 by John Wiley & Sons, Ltd, The Atrium, Southern Gate,

Chichester, West Sussex P019 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk

Reprinted March 2003, January 2004, December 2004, April 2006, February 2007, March 2008,

March 2010, March 2011

All Rights Reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except under the terms of fthe Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1T 4LP, UK, without the permission in writing of the Publisher, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the publication.

Neither the author(s) nor John Wiley & Sons, Ltd accept any responsibility or liability for loss or damage occasioned to any person or property through using the material, instructions, methods or ideas contained herein, or acting or refraining from acting as a result of such use. The author(s) and Publisher expressly disclaim all implied warranties, including merchantability of fitness for any particular purpose.

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons (Canada) Ltd, 22 Worcester Road, Etobicoke, Ontario M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Cataloging-in-Publication Data:

Grant, I. S. (Ian S.)

Electromagnetism / I. S. Grant, W. R. Phillips.—2nd ed.

p. cm.—(The Manchester physics series)

ISBN 978 0 471 92711 2—ISBN 978 0 471 92712 9 (pbk.)

1. Electromagnetism. I. Phillips, W. R. (William Robert)

II. Title. III. Series.

QC760.G76 1990

537—dc20 90–35818

CIP

British Library Cataloguing in Publication Data:

Grant, I. S.

Electromagnetism.—2nd ed.

1. Electromagnetism

I. Title II. Phillips, W. R. III. Series

537

ISBN 978 0 471 92711 2 (hbk)

ISBN 978 0 471 92712 9 (pbk)

Editors' preface to the Manchester Physics Series

The Manchester Physics Series is a series of textbooks at first degree level. It grew out of our experience at the Department of Physics and Astronomy at Manchester University, widely shared elsewhere, that many textbooks contain much more material than can be accommodated in a typical undergraduate course; and that this material is only rarely so arranged as to allow the definition of a shorter self-contained course. In planning these books we have had two objectives. One was to produce short books: so that lecturers should find them attractive for undergraduate courses; so that students should not be frightened off by their encyclopaedic size or their price. To achieve this, we have been very selective in the choice of topics, with the emphasis on the basic physics together with some instructive, stimulating and useful applications. Our second objective was to produce books which allow courses of different lengths and difficulty to be selected, with emphasis on different applications. To achieve such flexibility we have encouraged authors to use flow diagrams showing the logical connections between different chapters and to put some topics in starred sections. These cover more advanced and alternative material which is not required for the understanding of latter parts of each volume.

Although these books were conceived as a series, each of them is self-contained and can be used independently of the others. Several of them are suitable for wider use in other sciences. Each Author's Preface gives details about the level, prerequisites, etc., of his volume.

The Manchester Physics Series has been very successful with total sales of more than a quarter of a million copies. We are extremely grateful to the many students and colleagues, at Manchester and elsewhere, for helpful criticisms and stimulating comments. Our particular thanks go to the authors for all the work they have done, for the many new ideas they have contributed, and for discussing patiently, and often accepting, the suggestions of the editors.

Finally, we would like to thank our publishers, John Wiley & Sons Ltd, for their enthusiastic and continued commitment to the Manchester Physics Series.

D. J. Sandiford

F. Mandl

A. C. Phillips

February 1997

Preface to the Second Edition

The basic content of undergraduate courses in electromagnetism does not change rapidly, and the range of topics covered in the second edition of this book is almost the same as in the first edition. We have made a few additions, for example by giving a fuller treatment of circuit analysis and by discussing the dispersion of electromagnetic waves. Some material which now seems outdated has been removed, and illustrative examples have been modernized.

We have made many small changes in presentation which we hope will make the argument clearer to readers. We gratefully acknowledge the help of all those who have suggested ways of improving the text. We are particularly indebted to Dr. R. Mackintosh and his colleagues at the Open University for a host of detailed suggestions. The adoption of this book as the text for the new ‘third-level’ Open University course on electricity and magnetism led to a careful scrutiny of the first edition by the course team. This has resulted, we believe, in changes which make the book more useful for students. Any errors or obscurities which remain are our responsibility.

Manchester

January, 1990

I. S. GRANT

W. R. PHILLIPS

Preface to the First Edition

This book is based on lectures on classical electromagnetism given at Manchester University. The level of difficulty is suitable for honours physics students at a British University or physics majors at an American University. A-level or high school physics and calculus are assumed, and the reader is expected to have some elementary knowledge of vectors. Electromagnetism is often one of the first branches of physics in which students find that they really need to make use of vector calculus. Until one is used to them, vectors are difficult, and we have accordingly treated them rather cautiously to begin with. Brief descriptions of the properties of the differential vector operators are given at their first appearance. These descriptions are not intended to be a substitute for a proper mathematical text, but to remind the reader what div, grad and curl are all about, and to set them in the context of electromagnetism. The distinction between macroscopic and microscopic electric and magnetic fields is fully discussed at an early stage in the book. It is our experience that students do get confused about the fields E and D, or B and H. We think that the best way to help them overcome their difficulties is to give a proper explanation of the origin of these fields in terms of microscopic charge distributions or circulating currents.

The logical arrangement of the chapters is summarized in a flow diagram on the inside of the front cover. Provided that one is prepared to accept Kirchhoff's rules and the expressions for the e.m.f.s across components before discussing the laws on which they are based, the A.C. theory in Chapters 7 and 8 does not require any prior knowledge of the earlier chapters. Chapters 7 and 8 can therefore be used at the beginning of a course on electromagnetism. Sections of the book which are starred may be omitted at a first reading, since they do not contain material needed in order to understand later chapters.

We should like to thank the many colleagues and students who have helped with suggestions and criticisms during the preparation of this book; any errors which remain are our own responsibility. It is also a pleasure to thank Mrs Margaret

King and Miss Elizabeth Rich for their rapid and accurate typing of the manuscript.

May, 1974

Manchester, England.

I. S. GRANT

W. R. PHILLIPS

CHAPTER 1

Force and energy in electrostatics

The only laws of force which are known with great precision are the two laws describing the gravitational forces between different masses and the electrical forces between different charges. When two masses or two charges are stationary, then in either case the force between them is inversely proportional to the square of their separation. These inverse square laws were discovered long ago: Newton's law of gravitation was proposed in 1665, and Coulomb's law of electrostatics in 1785. This chapter is concerned with the application of Coulomb's law to systems containing any number of stationary charges. Before studying this topic in detail, it is worth pausing for a moment to consider the consequences of the law in the whole of physics.

In order to make full use of our knowledge of a law of force, we must have a theory of mechanics, that is to say, a theory which describes the behaviour of an object under the action of a known force. Large objects which are moving at speeds small compared to the speed of light obey very closely the laws of classical Newtonian mechanics. For example, these laws and the gravitational force law together lead to accurate predictions of planetary motion. But classical mechanics does not apply at all to observations made on particles of atomic scale or on very fast-moving objects. Their behaviour can only be understood in terms of the ideas of quantum theory and of the special theory of relativity. These two theories have changed the framework of discussion in physics, and have made possible the spectacular advances of the twentieth century.

It is remarkable that while mechanics has undergone drastic amendment, Coulomb's law has stood unchanged. Although the behaviour of atoms does not fit the framework of the old mechanics, when the Coulomb force is used with the

theories of relativity and quantum mechanics, atomic interactions are explained with great precision in every instance when an accurate comparison has been made between experiment and theory. In principle, atomic physics and solid state physics, and for that matter the whole of chemistry, can be derived from Coulomb’s law. It is not feasible to derive everything in this way, but it should be borne in mind that atoms make up the world around us, and that its rich variety and complexity are governed by electrical forces.

1.1 ELECTRIC CHARGE

Most of this book applies electromagnetism to large-scale objects, where the atomic origin of the electrical forces is not immediately apparent. However, to emphasize this origin, we shall begin by consideration of atomic systems. The simplest atom of all is the hydrogen atom, which consists of a single proton with a single electron moving around it. The hydrogen atom is stable because the proton and the electron attract one another. In contrast, two electrons repel one another, and tend to fly apart, and similarly the force between two protons is repulsive*. These phenomena are described by saying that there are two different kinds of electric charge, and that like charges repel one another, whereas unlike charges are attracted together. The charge carried by the proton is called positive, and the charge carried by the electron negative.

The magnitude and direction of the force between two stationary particles, each carrying electric charge, is given by Coulomb's law. The law summarizes four facts:

(i) Like charges repel, unlike charges attract.

(ii) The force acts along the line joining the two particles.

(iii) The force is proportional to the magnitude of each charge.

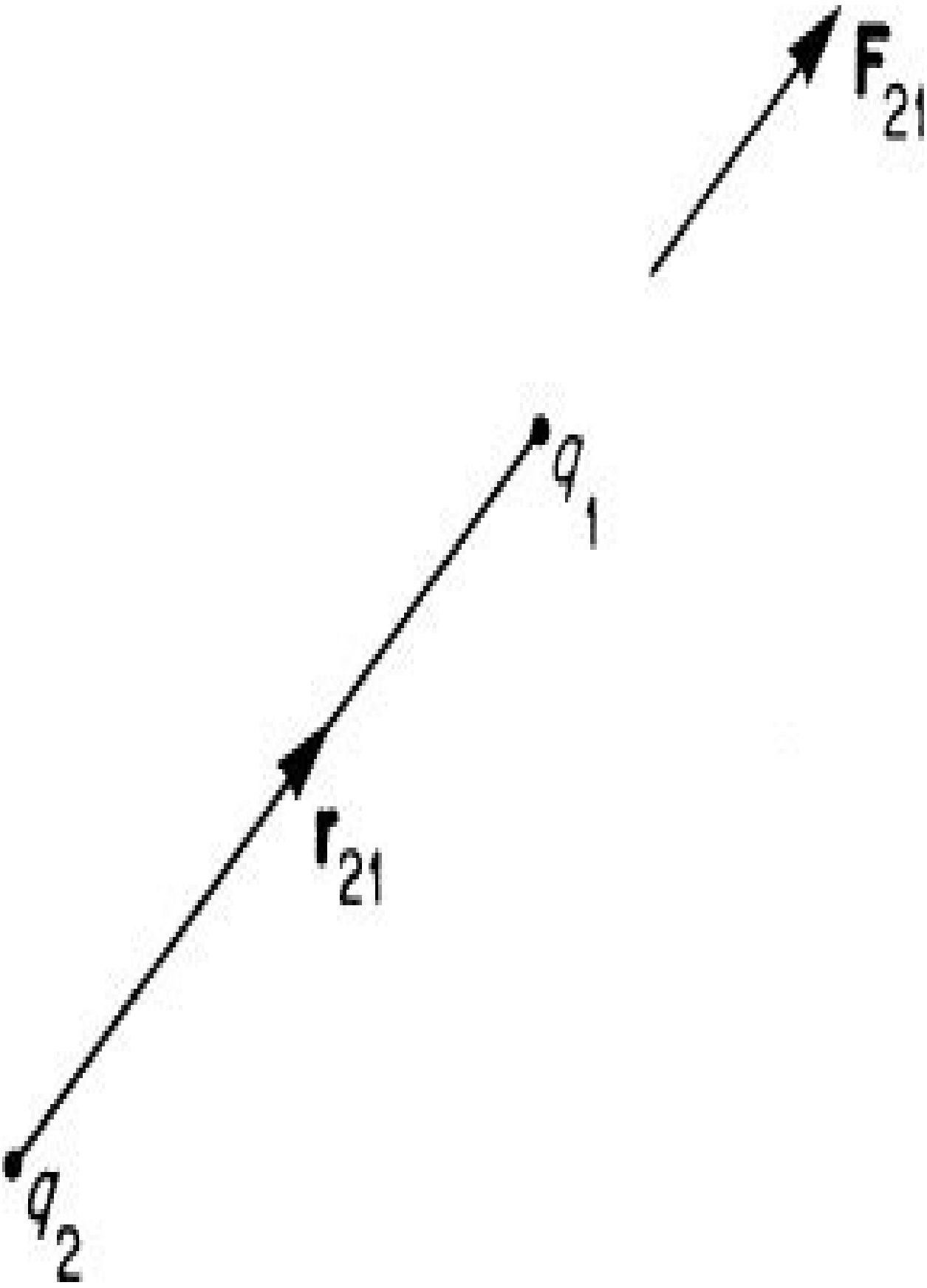
(iv) The force is inversely proportional to the square of the distance between the particles.

The mathematical statement of Coulomb's law is:

(1.1)

The vector F21 in Figure 1.1 represents the force on particle 1 (carrying a charge q1) exerted by the particle 2 (carrying charge q2). The line from q2 to q1 is represented by the vector r21, of length r21: since the unit vector along the direction r21 can be written r21/r21, Equation (1.1) is an inverse square law of force, although appears in the denominator. Notice that the equation automatically accounts for the attractive or repulsive character of the force if q1 and q2 include the sign of the charge. When the charges q1 and q2 are both positive or both negative, the force on q1 is along r21, i.e. it is repulsive. On the other hand, when one charge is positive and the other negative, the force is in the direction opposite to r21, i.e. it is attractive.

Figure 1.1. The force between two charges.


F_{21}
q_1
r_{21}
q_2

To complete the statement of the force law, we must decide what units to use, and hence determine the constant of proportionality in Equation (1.1). We shall use SI (Systeme International) units, which are favoured by most physicists and engineers applying electromagnetism to problems involving large-scale objects. A different system of units, called the Gaussian system, is frequently used in atomic physics and solid state physics, and it is an unfortunate necessity for students to become reasonably familiar with both systems. (The two systems of units are discussed in Appendix A.) In SI units, Coulomb's law is written as

(1.2)

where

q1 and q2 are measured in coulombs,

r21 is measured in metres

and

F21 is measured in newtons.

The magnitude of the unit of charge, which is called the coulomb, is actually defined in terms of magnetic forces, and we shall leave discussion of the definition until Chapter 4. The factor 4π in the constant of proportionality in Coulomb's law is introduced in order to simplify some important equations

which we shall meet later. The constant ε_0, which is called the permittivity of free space, has the value

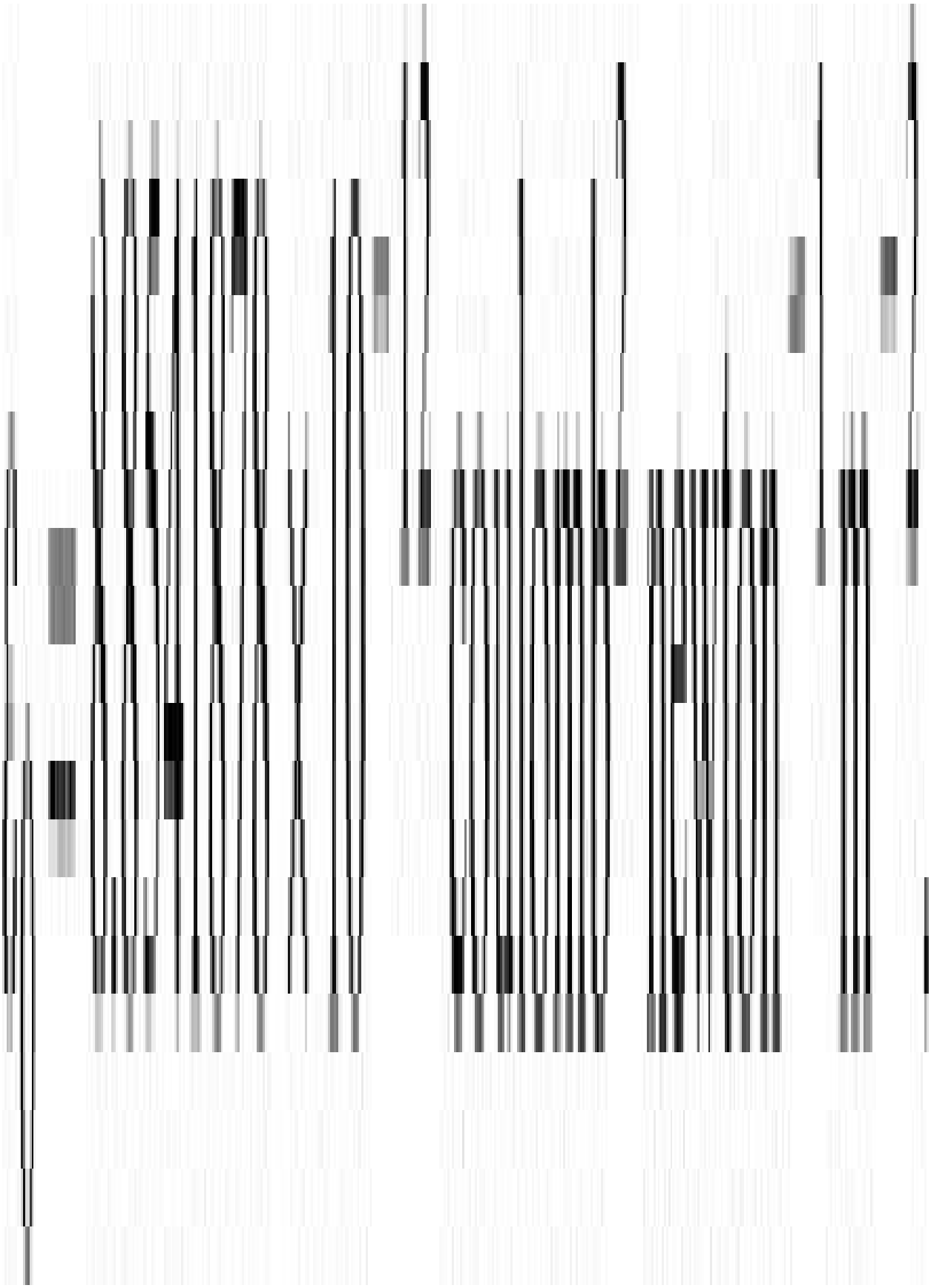
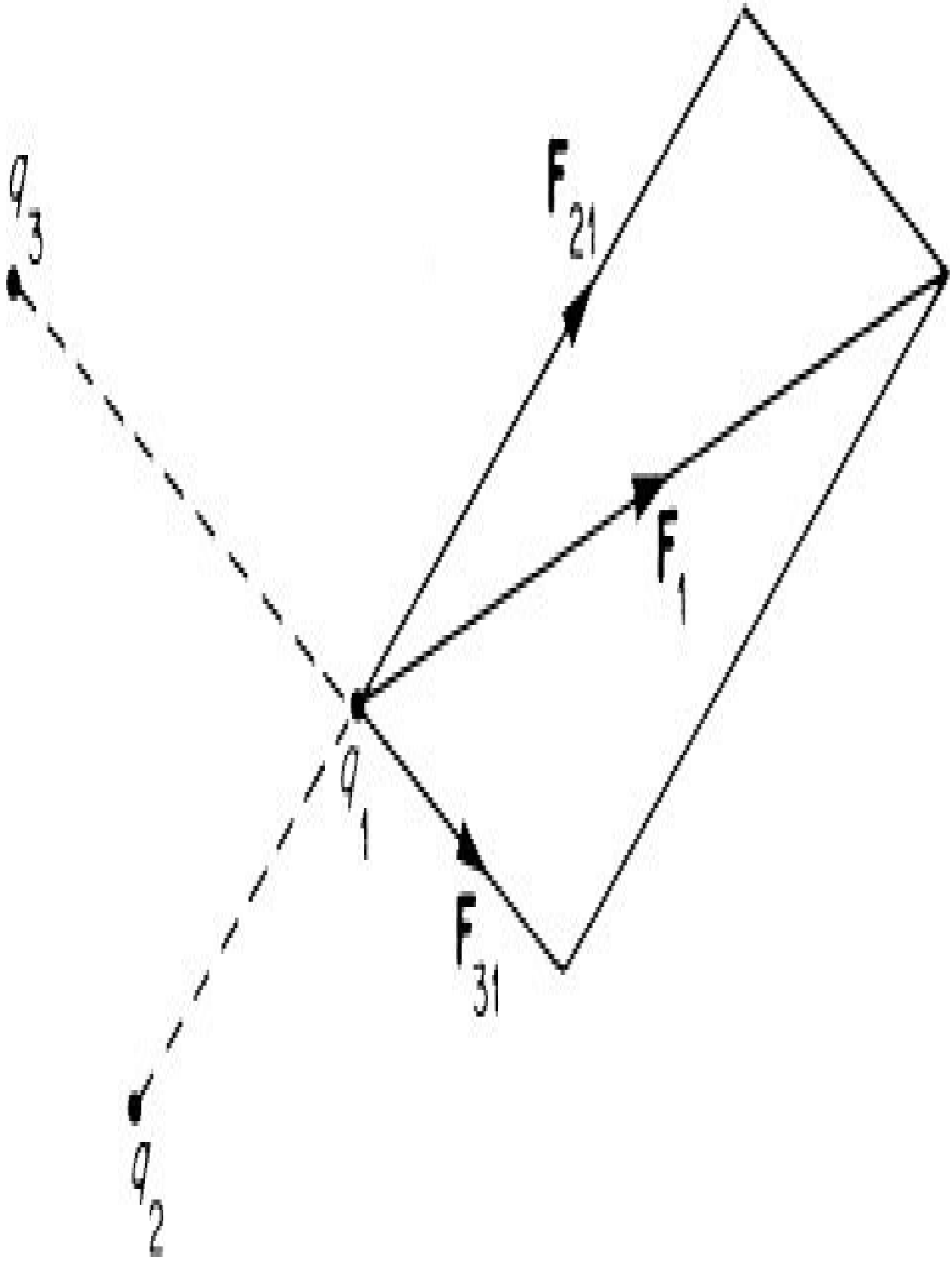
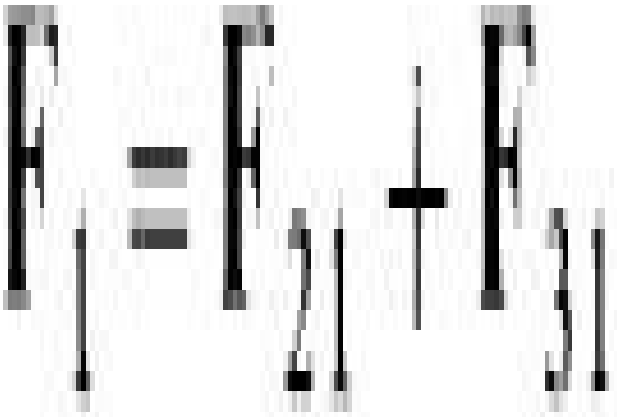


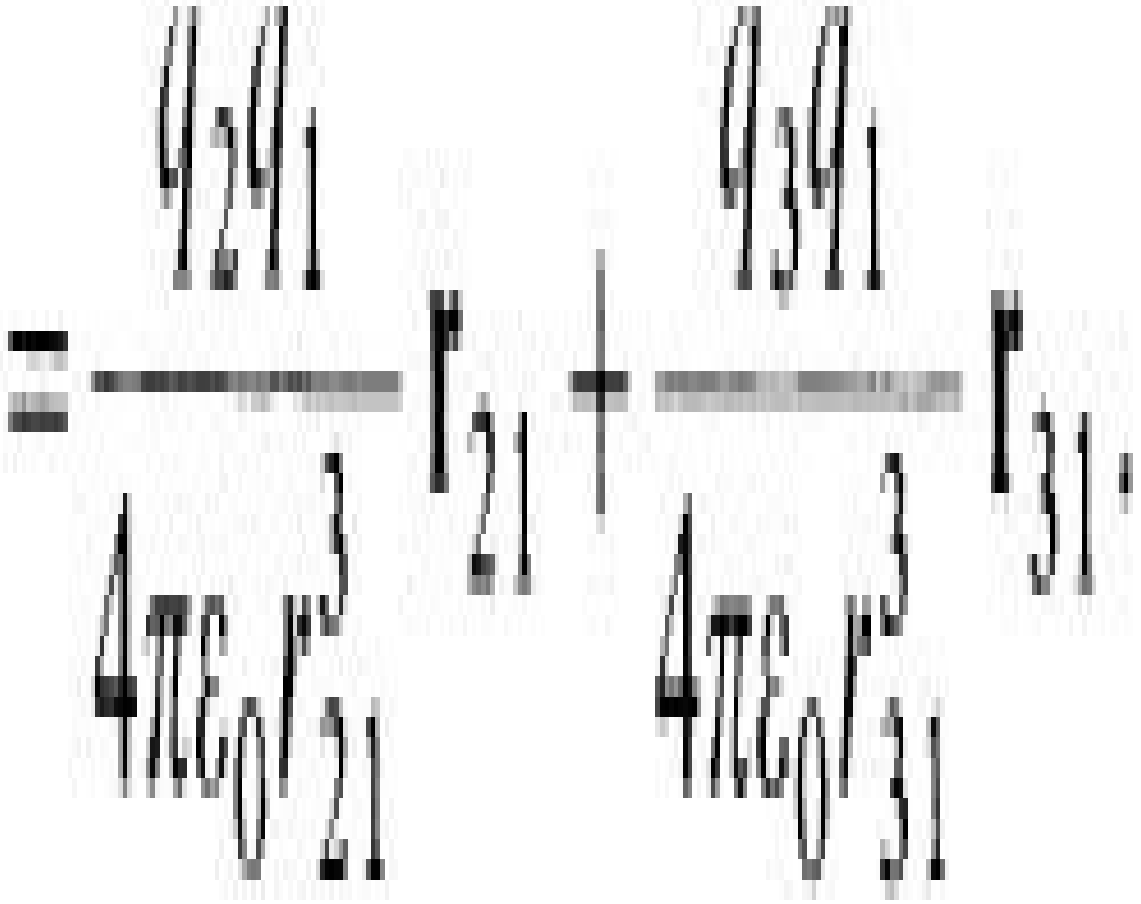
Figure 1.2. How electrostatic forces are added when there are more than two charges.


q3
F21
F1
q1
F31
q2

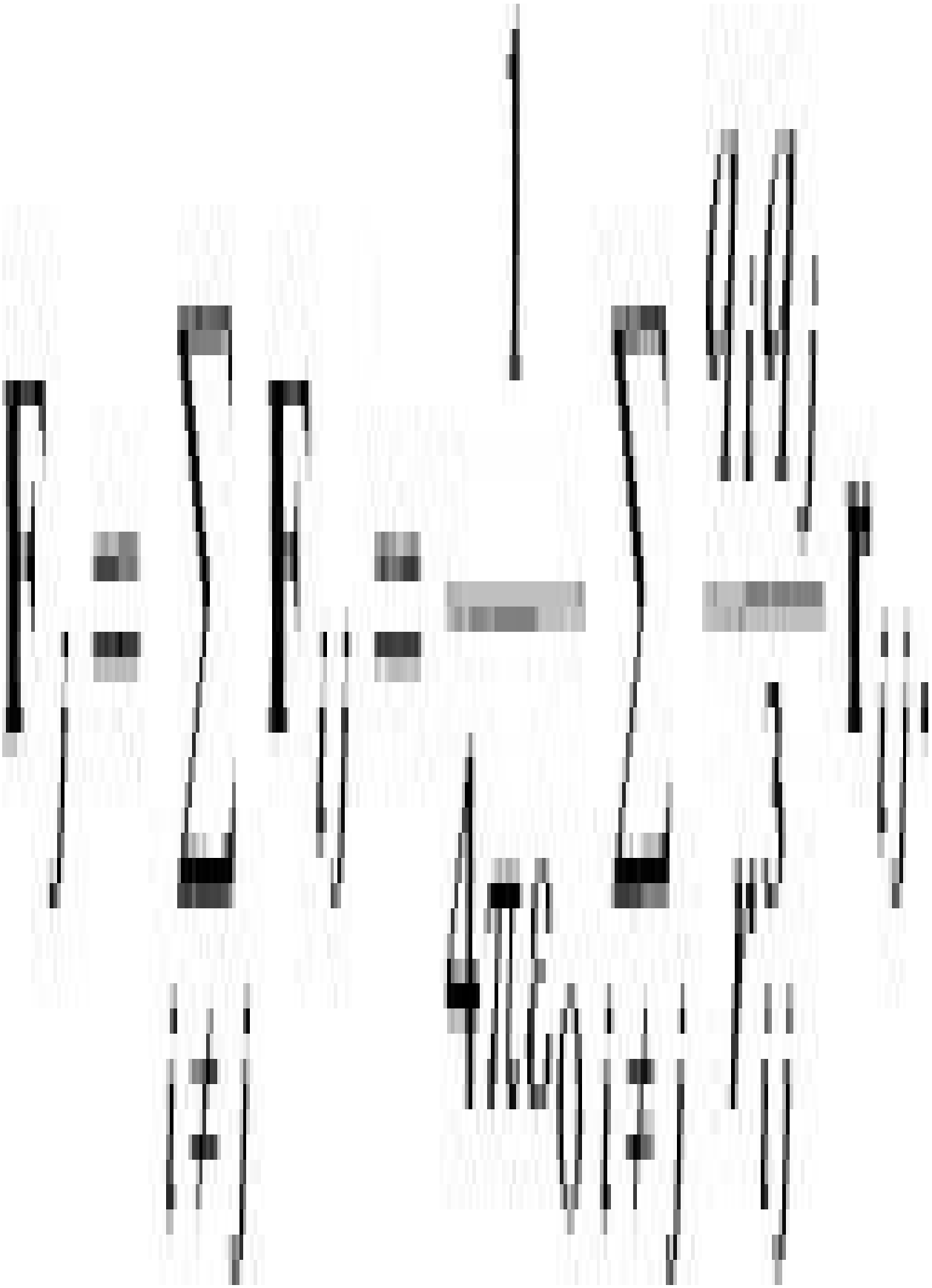
The value of ε0 is not determined experimentally, but has been defined in a way which makes the SI system of units self-consistent: the relation between electrical units and other units is discussed in Appendix A.

Electrostatic forces are two-body forces, which means that the force between any pair of charges is unaltered by the presence of other charges in their neighbourhood*. In a system containing many charges, the electrostatic force between each pair is given by Coulomb's law. To find the total force on any one particle, one simply makes a vector sum of the forces it experiences due to all the others separately. This rule is illustrated in Figure 1.2 for a system of three charges. The forces on q1 due to the presence of the charges q2 and q3 are F21 and F31 and the total force on q1 is





In general, the force Fj on a charge qj due to a number of other charges qi is



The symbol $i \neq j$ under the summation signs indicates that the summation for charge i is over all the other charges j, but of course not including charge i itself. This equation can be written in another way in terms of the position vectors of the charges with respect to a fixed origin O. If the position vectors of the charges q1, q2 … qi … are r1, r2 … ri … then the vector joining charges i and j is rij = rj – ri. The total force on qj is thus

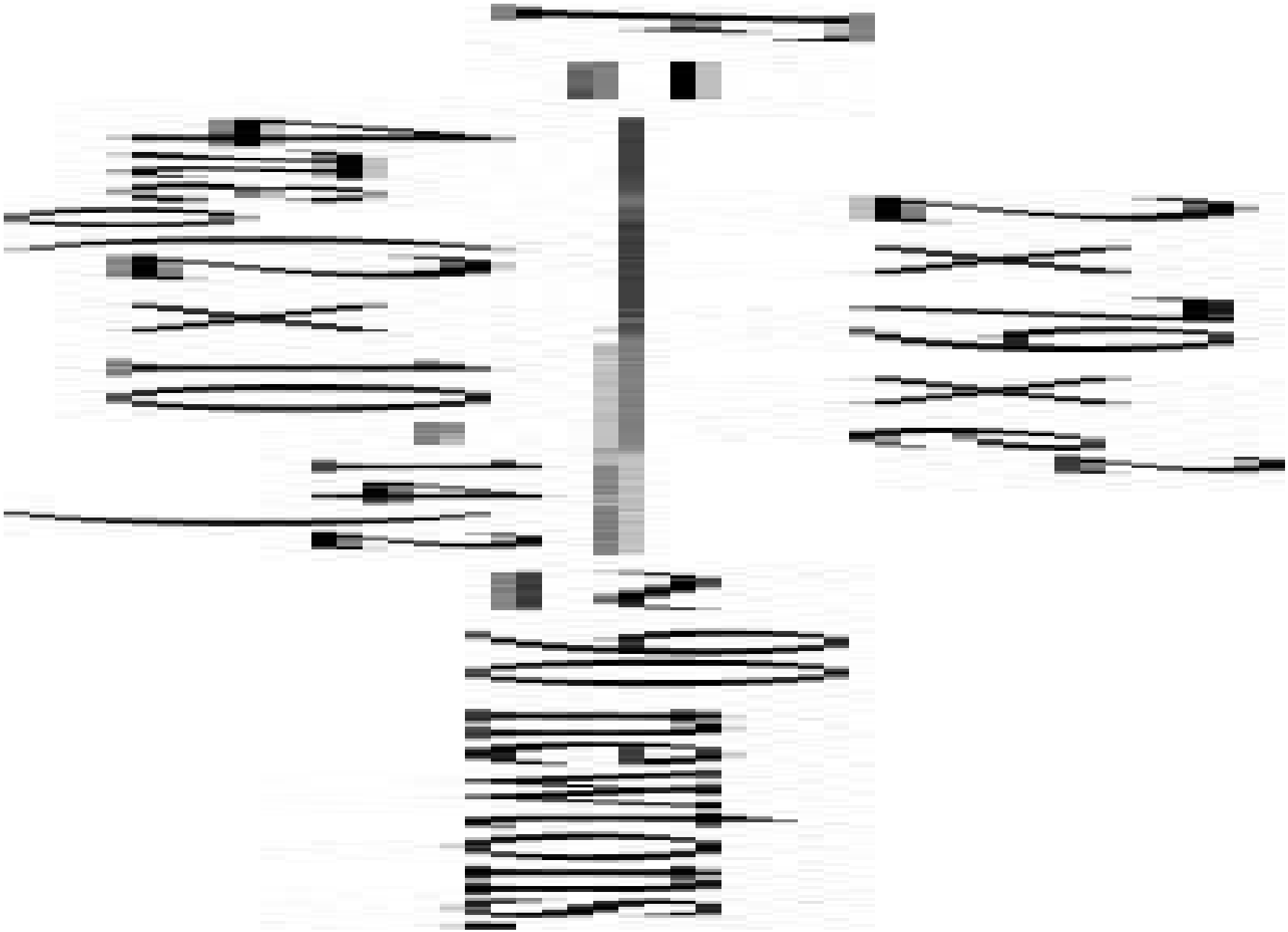
(1.3)

A trivial example of the application of Equation (1.3) is in working out the electrostatic forces exerted by atomic nuclei containing many protons on the electrons surrounding them. Nuclei are much smaller than atoms, and for this purpose can be regarded as point charges. Equation (1.3) then tells us that the attractive force between an electron and a nucleus containing Z protons is Z times as great as that between an electron and a single proton.

It turns out that apart from the sign, the charge carried by electrons and protons is the same, and has the magnitude $e = 1.602 \times 10^{-19}$ coulombs*: the charge on the proton is +e, that on the electron is – e. The strength of atomic interactions is governed by the size of the electronic charge e. Although e is a very small number when expressed in coulombs, this does not imply that electrostatic forces are feeble. On the contrary, they are immensely strong. For example, electrostatic forces are responsible for the great strength of solids under compression. When neighbouring atoms are close together, their electron clouds begin to overlap, and the mutual repulsion of these clouds opposes any compressing force.

Another example of the strength of the electrostatic force acting on the atomic scale is given by the experiment which led Rutherford to propose the nuclear model of the atom. He found that when swiftly moving α-particles are allowed to

collide with gold atoms, they are sometimes deflected through 180°, implying that a strong force is at work. The force is just the electrostatic repulsion experienced by an α-particle when it chances to approach close to the nucleus of a gold atom. Let us calculate the magnitude of the force. An α-particle is a helium nucleus, containing two protons and carrying a charge + 2e, and the gold nucleus carries a charge + 79e. In Rutherford's experiment, the α-particles were energetic enough to approach within 2×10^{-14} m of the nucleus (still well outside the range of the nuclear forces). Substituting in Equation (1.3), the repulsive force at this distance is



This force, acting within a single atom, is nearly as much as the weight of a mass of 10 kilogrammes!

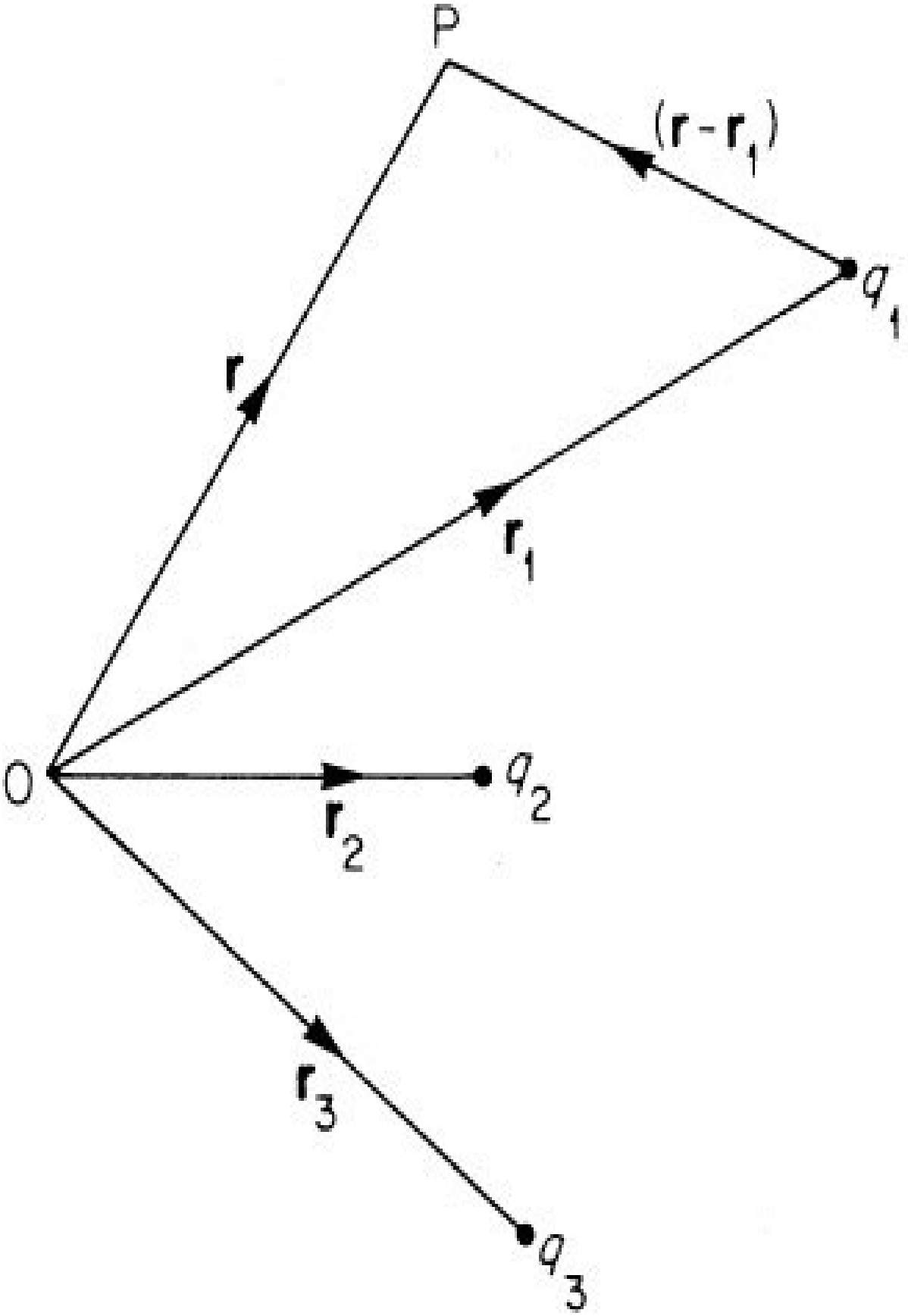
Normally we do not notice that electrostatic forces are so powerful, because matter is usually electrically neutral, carrying equal amounts of positive and negative charge. Not only are large lumps of matter electrically neutral, but the positive charge on the nucleus of a single isolated atom is precisely cancelled out by the negative charge of the surrounding electrons. So far as we know the cancellation is exact, and it has been shown experimentally that the magnitude of the net residual charge on a neutral atom is less than $10^{-20}e$. This is very remarkable, since apart from their electrical behaviour, protons and electrons are totally dissimilar particles. Many charged particles besides electrons and protons have been discovered by nuclear physicists, and all* share the property of carrying charges $\pm e$. It follows that the total charge carried by any piece of matter must be an integral multiple of the electronic charge e. A situation like this one, in which a physical quantity is not allowed to have a continuous range of values, but is restricted to a set of definite discrete values, is referred to as a quantum phenomenon. No one knows why electric charge should obey this quantum rule; it is an experimental fact. Nevertheless, because the rule is universal in its application, we can be sure that the electronic charge is a physical quantity of fundamental importance.

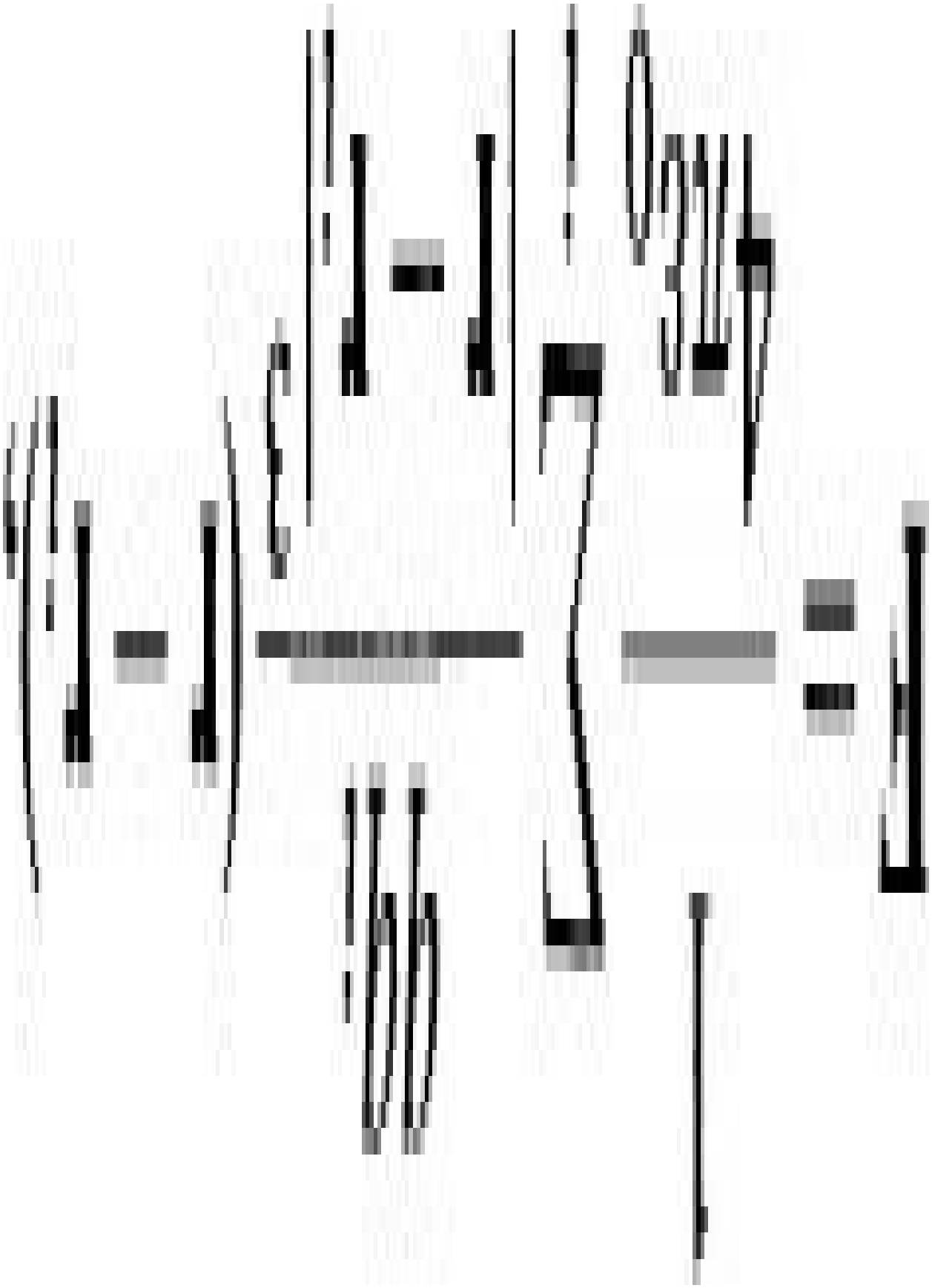
1.2 THE ELECTRIC FIELD

We have seen that the total force on a charged particle is the vector sum of the forces exerted on it by all other charges. Usually there is an enormous number of charged particles present in real matter. When considering the forces acting on any one of them, it is helpful to distract attention from the multitude of sources contributing to the net force by introducing the concept of the electric field. If a charge q experiences a force F, then the ratio F/q is called the electric field at the point where q is located. The dimensions of electric field are [force] [charge]−1, and in SI units the electric field is measured in newton coulomb−1. (An equivalent unit, which will be explained when we come to deal with electrostatic energy, is the volt m−1; 1 volt m−1 ≡ 1 newton coulomb−1.)

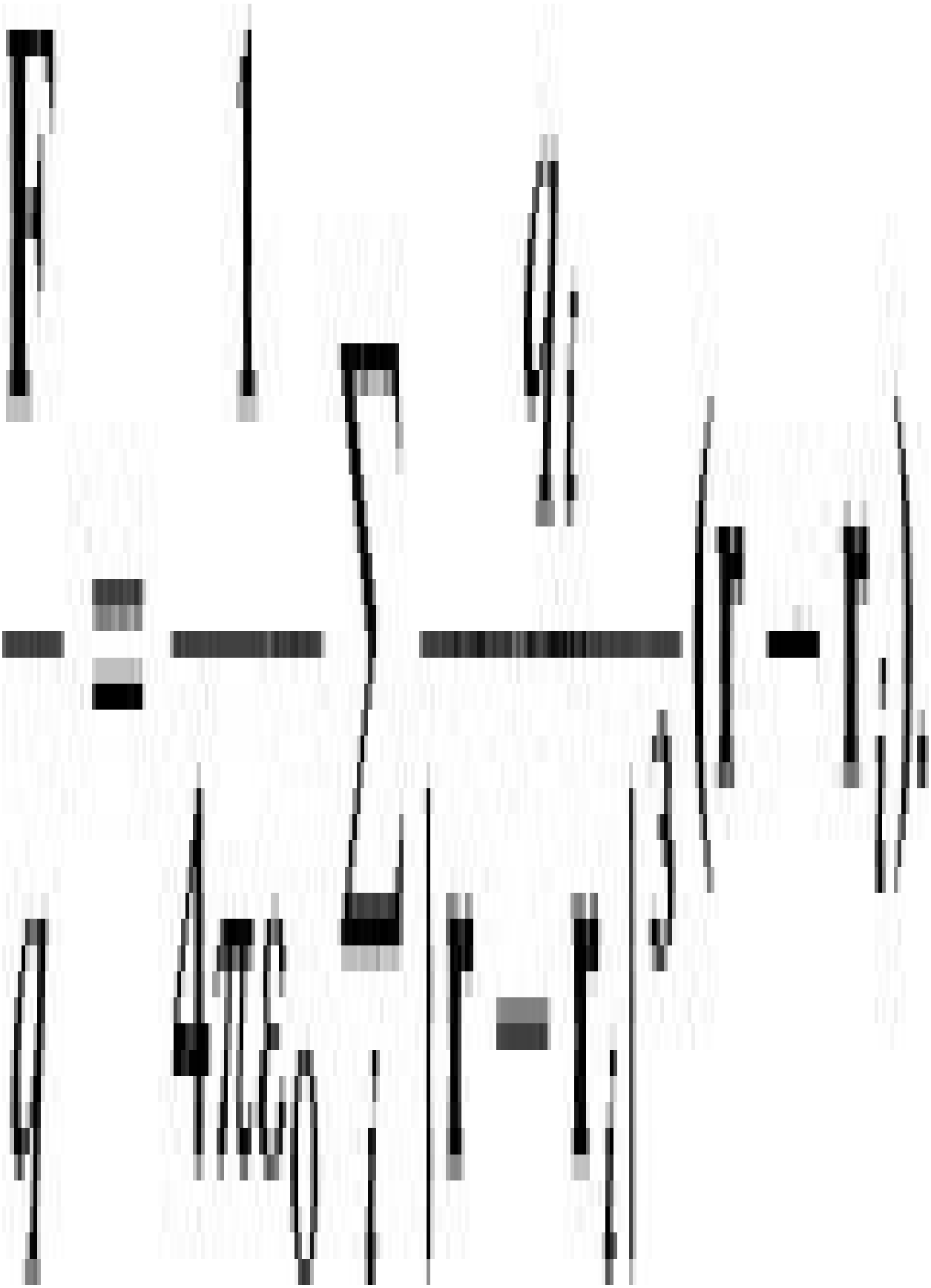
The electric field acting on q can be expressed in terms of the magnitudes of the other charges in the neighbourhood of q and their relative positions with respect to q. Let us assume that q is a test charge which can be put anywhere, and that its magnitude is very small, so that it exerts negligible forces on the other charges and can be moved about without altering their positions. Now we can evaluate the electric field at a point P, caused by an assembly of charges qi by placing the test charge at P. In Figure 1.3 P has a position vector r, and the charges qi have position vectors rf with respect to an origin at O. The force on the test charge is given by Equation (1.3) as

Figure 1.3. Vectors used in the definition of the electric field.


P
$(\mathbf{r}-\mathbf{r}_1)$
q_1
$\mathbf{r}$
$\mathbf{r}_1$
O
q_2
$\mathbf{r}_2$
$\mathbf{r}_3$
q_3

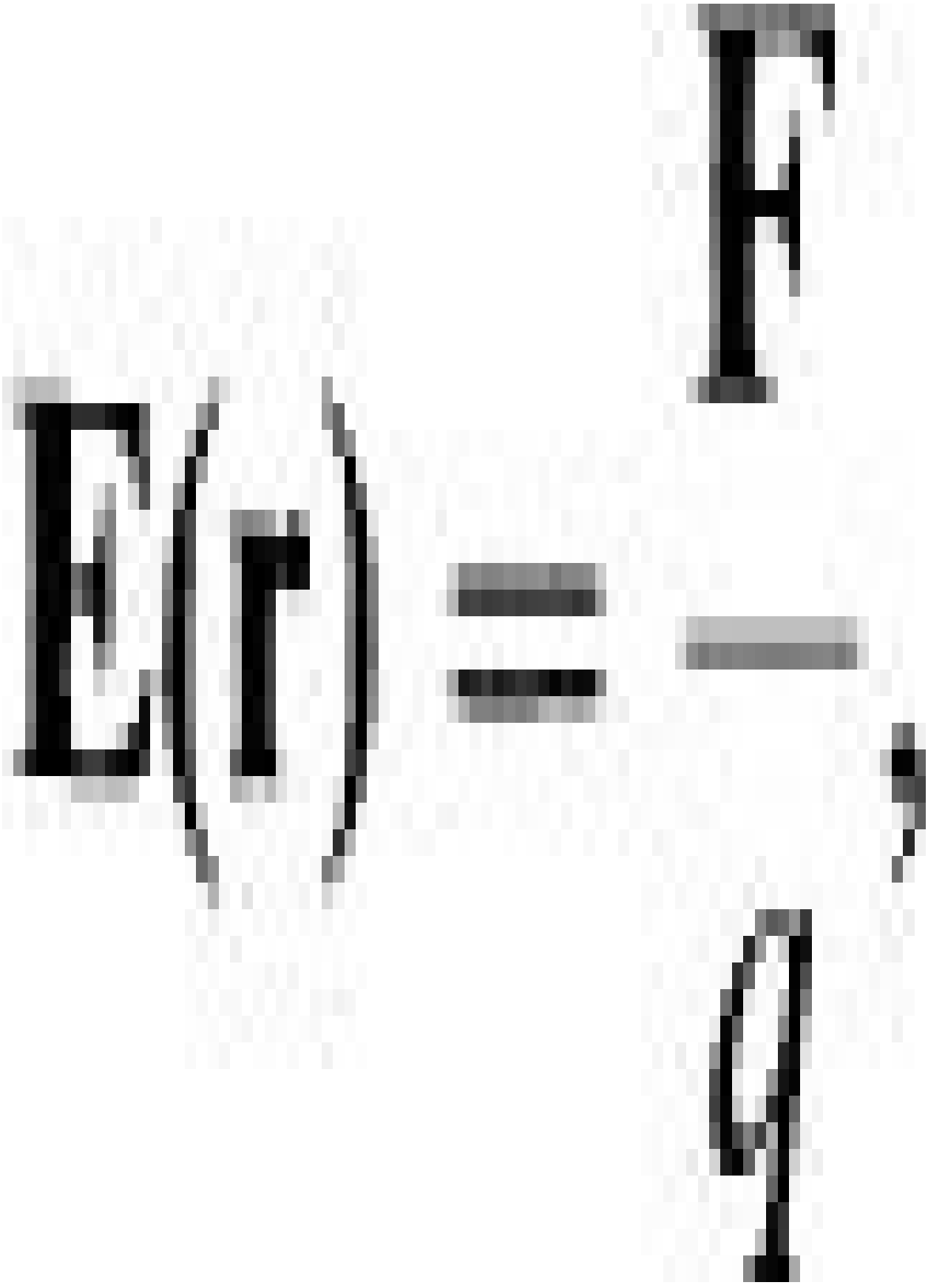


where (r – ri) represents the vector joining qi to the point P. The factor q is common to all terms in the sum, and it follows that



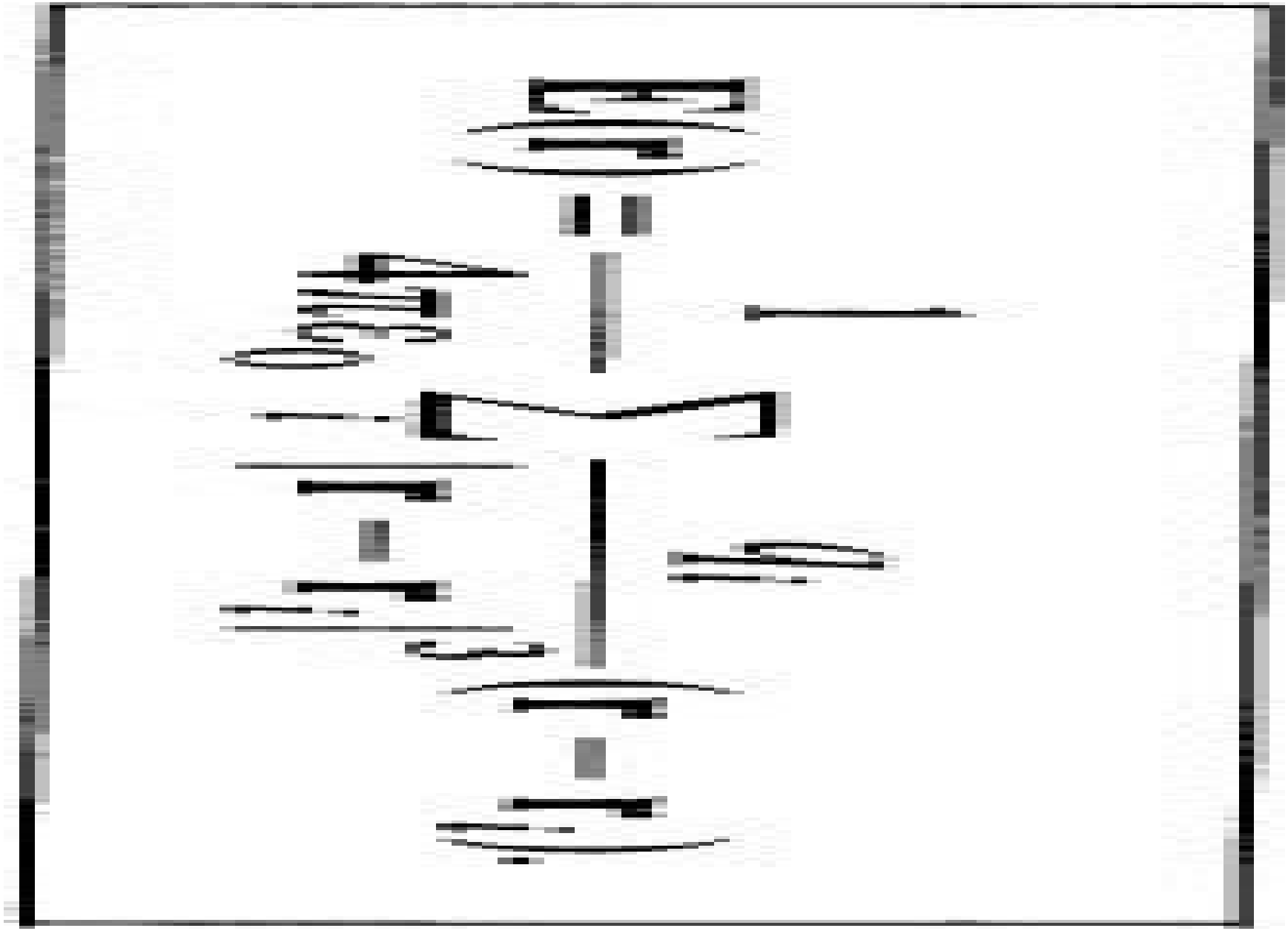
The magnitude q of the test charge does not appear on the right-hand side of this equation. We can therefore allow q to become vanishingly small—then we are quite sure that the presence of the test charge does not modify the position of the other charges. Let us call the electric field at the point P with position vector r E(r).

Then



or

(1.4)



Equation (1.4) is the definition of the electric field E(r), and it contains no reference to a test charge q. The test charge was introduced because it illustrates that the electric field is a force per unit charge, and because it helps one to visualize the electric field if one imagines a test charge which can be moved around to sample the strength of the field at any position. The electric field E(r) is a function of position; just as the value of a function f(x) is determined by the argument x, so the value of E(r) is given by Equation (1.4) in terms of its argument, which is the position vector r. The function E(r) is itself a vector, specifying the direction as well as the magnitude of the force per unit charge on a point charge at r. The electric field is only the first of a number of functions of position which are useful in electromagnetism. Those functions of position which are themselves vectors are called vector fields. When discussing vector fields we shall frequently omit any reference to the argument, writing the electric field, for example, simply as 'E', leaving it to be understood that the field is a function of position. Whenever there is some uncertainty about the position at which the function is to be evaluated, we shall always write out the vector field in full, including the argument.

In Equation (1.4) each charge qi appears just once on the right-hand side. If the charge qi were the only charge present, the field would be

(1.5)

and we can rewrite Equation (1.4) as

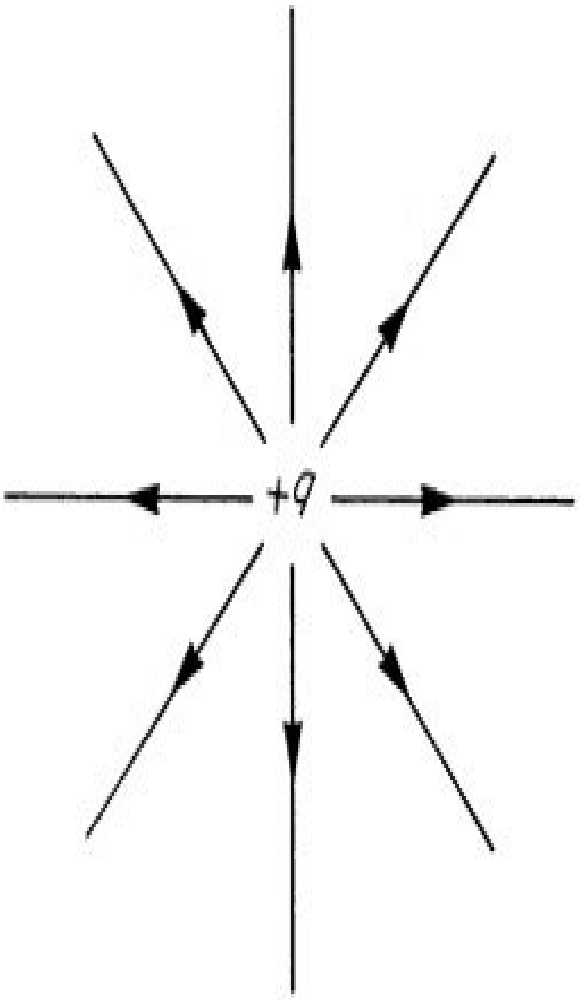
(1.6)

In other words, the total electric field is the sum of the electric fields due to each

charge separately. This is an example of the Principle of Superposition. As applied to electrostatic fields, the Principle of Superposition states that if we have two electrostatic fields due to different groups of charged particles, the fields must be added together (superimposed) to find the field due to all the particles in both groups. This follows immediately from Equation (1.6). We shall find that the Principle of Superposition applies to all the different kinds of field which occur in electromagnetism.

Now let us investigate the properties of the electric field in the neighbourhood of isolated point charges. From Equation (1.5), the magnitude of the field at a distance r from a positive point charge +q is $q/4\pi\varepsilon_0 r^2$, and the field points away from the charge. The field around a negative point charge – q has the same magnitude, but it points towards the charge. In Figure 1.4 the direction of the field around positive and negative charges is indicated by the arrowed lines. These continuous lines, everywhere following the direction of the field, are called lines of force or field lines. Lines of force begin on positive charges and end on negative charges, but they may also go to infinity without terminating, as in Figure 1.4. Notice that the lines of force are close together near the point charges where the field is strong, and far apart at large distances where the field is weak.

Figure 1.4 Field lines around point charges.


+q

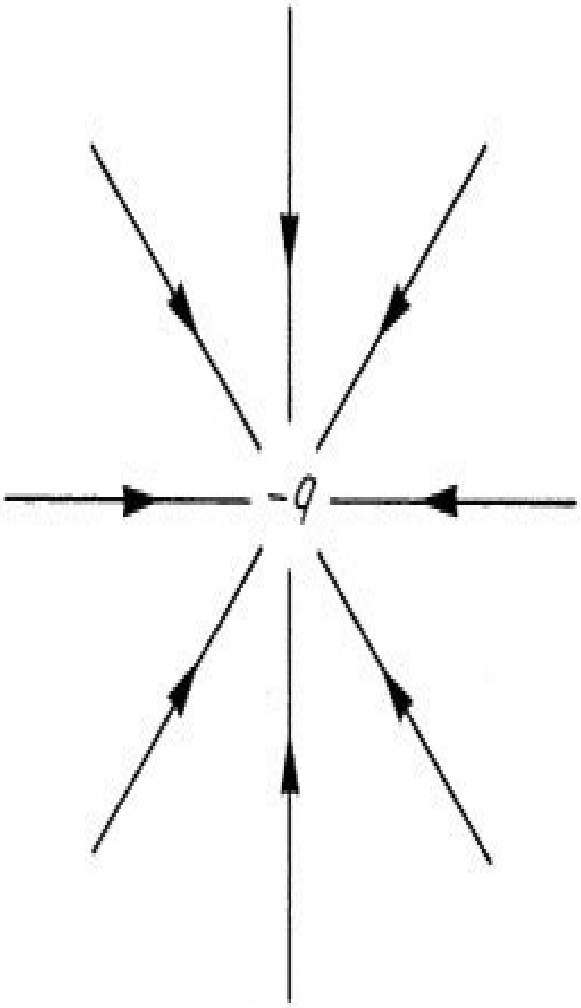
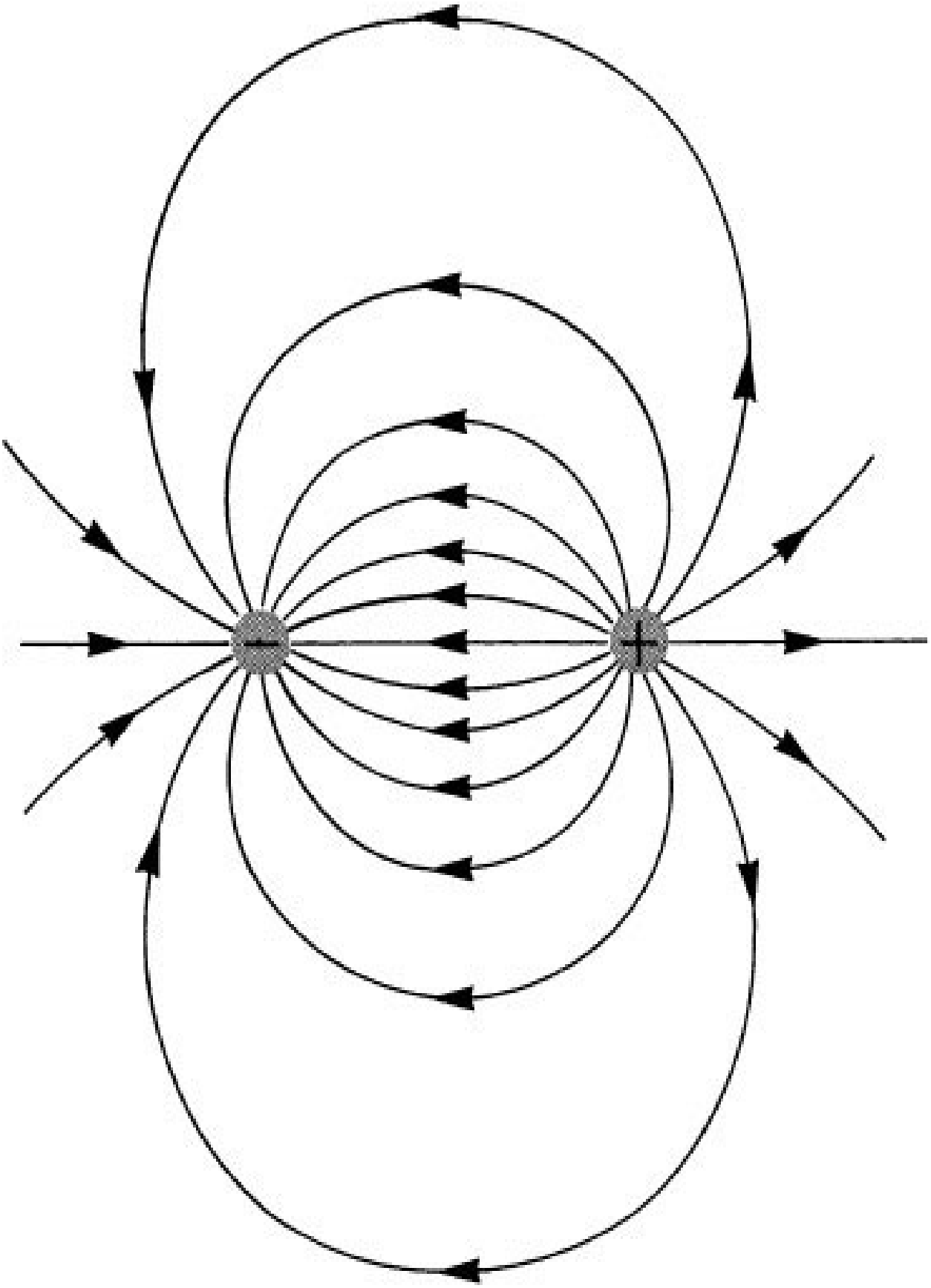

-q

Figure 1.5 Field lines around an electric dipole.

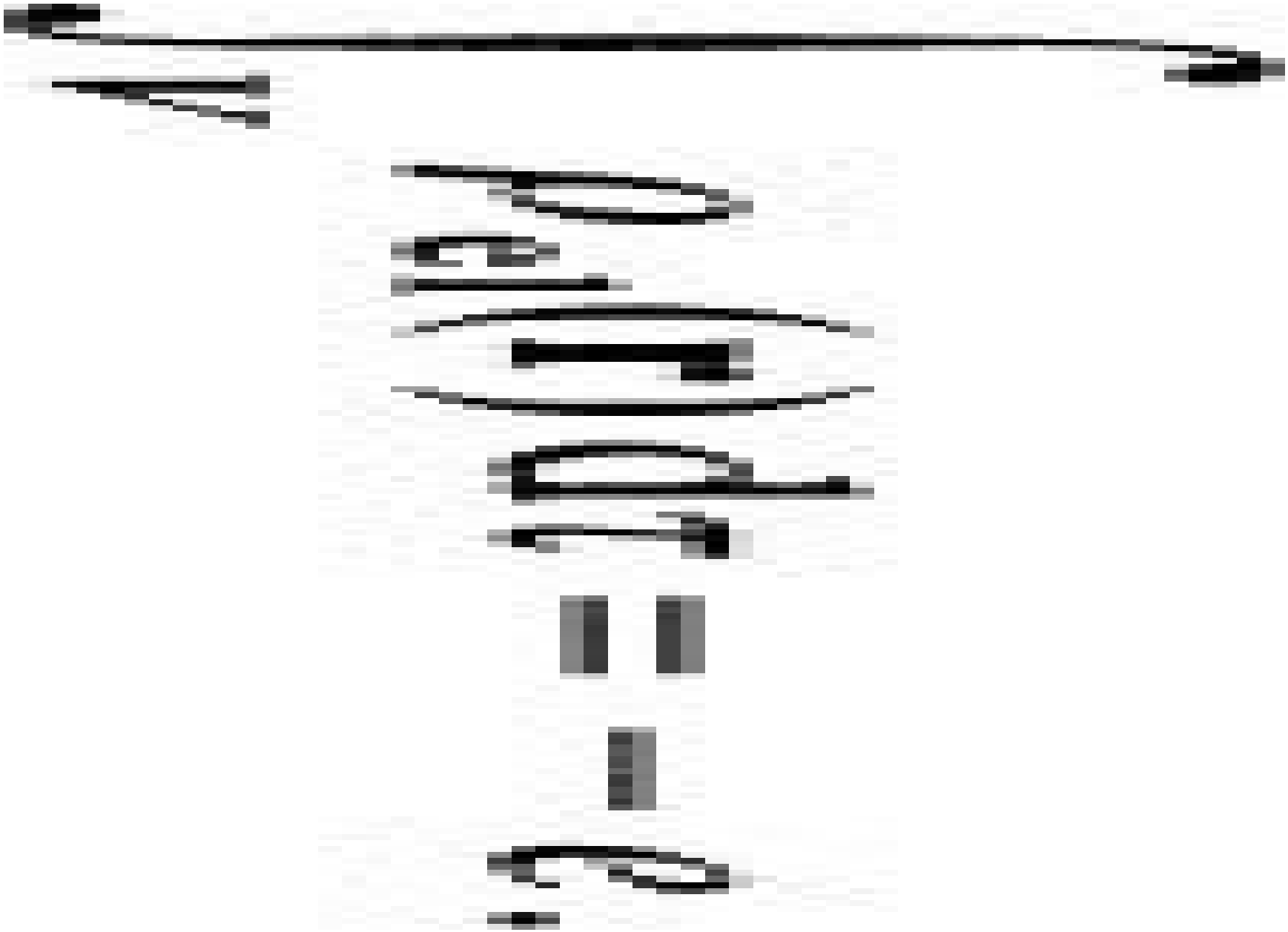


We can also draw diagrams of lines of force to illustrate the electric field when there are many charges present. Lines of force are continuous, except where they terminate on positive or negative charges, and they never cross one another, since the direction of the field is unique at every point. One can often get a rough idea of the field around a distribution of charges simply by sketching lines of force, and without doing any mathematics. For example, Figure 1.5 shows the lines of force near a pair of point charges of equal magnitude, one positive and one negative. Such a pair of equal and opposite charges is called an electric dipole. Very close to each charge, the field is almost the same as for isolated point charges, but the field lines starting off at the positive charge curve round to finish at the negative charge. The diagram gives an indication of the strength of the field as well as its direction, because lines of force are always densely packed in regions where the field is strong. The total number of lines on the diagram is not significant; if twice as many lines were drawn terminating on each charge, regions of relatively high or low density of lines of force would still correspond to regions of high or low field strength. The example of the electric dipole is an important one and later on we shall obtain a mathematical expression for the field at some distance from a dipole.

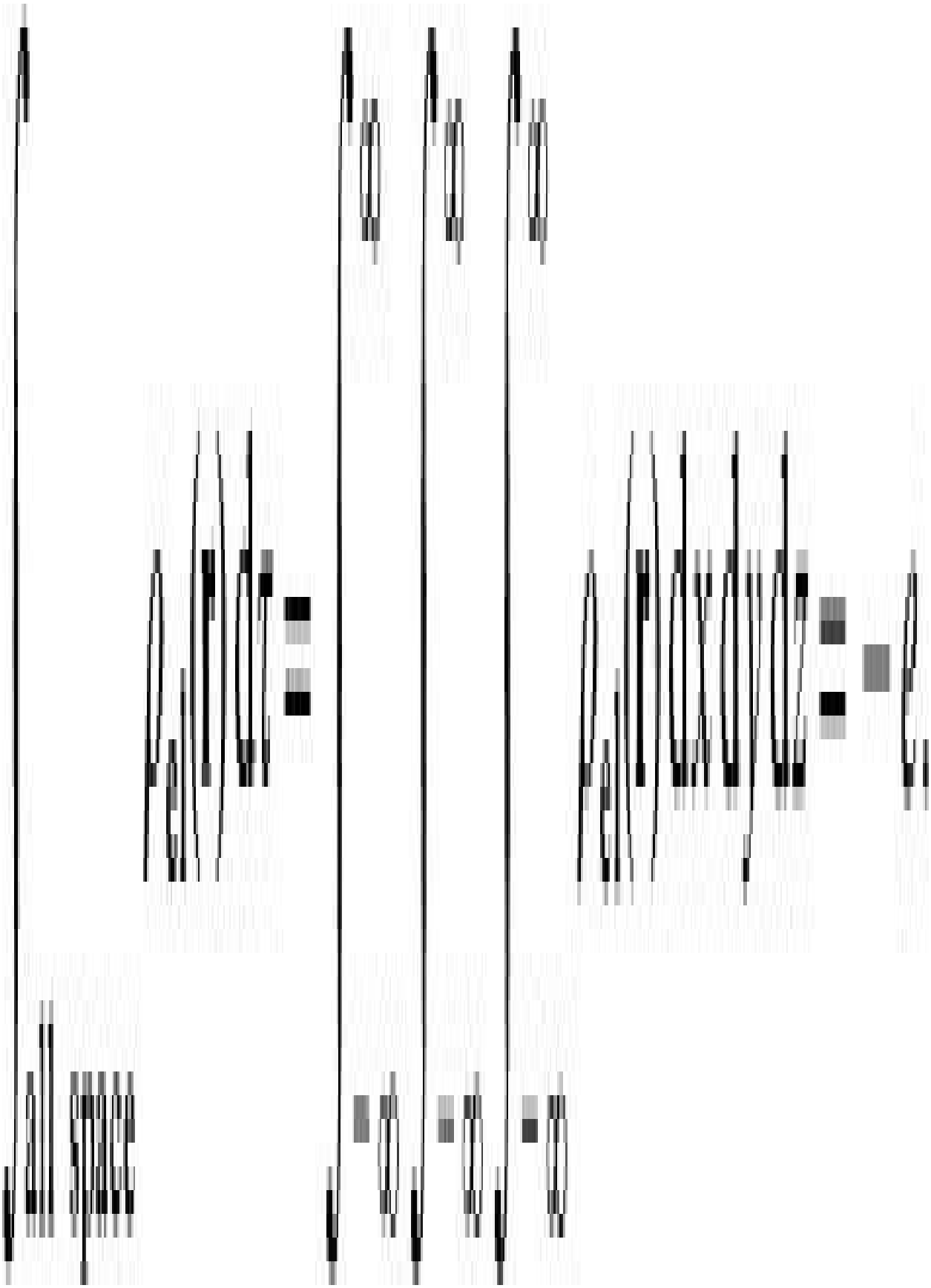
1.3 ELECTRIC FIELDS IN MATTER

1.3.1 The atomic charge density

So far electric fields have been dealt with in terms of idealized particles which are stationary and which carry point charges. This is not satisfactory if we want to discuss the electric field within an actual atom, since electrons are certainly not stationary point charges. Indeed, according to quantum mechanics the position of an electron cannot even be sharply defined. Instead of imagining the electron as a point, it is more realistic to regard its charge as being smeared out in a cloud around the nucleus. The electric field in and around a hydrogen atom, for example, behaves as though only part of the electronic charge is contained by any section of the electron cloud. Let us write $\rho_{el}(r)\delta\tau$ for the charge contained within a small volume $\delta\tau$ situated at the point with position vector r. The quantity $\rho_{el}(r)$ is the charge density of the electron cloud, i.e. its charge per unit volume at r. The charge density is another function of position, but unlike the electric field it is a scalar quantity, fully specified by its magnitude at each point. Scalar functions of position such as the charge density are called scalar fields. As with the vector fields, we shall often omit the argument of scalar fields, writing the electron charge density just as ρ_{el}, for example. The total charge $(-e)$ carried by the electron in a hydrogen atom is equal to the sum of the charges contained in all the small volumes $\delta\tau$ making up the electron cloud. In the limit as the volumes $\delta\tau$ become infinitesimal, the sum becomes an integral; integrating over a volume V large enough to contain the whole atom, we write



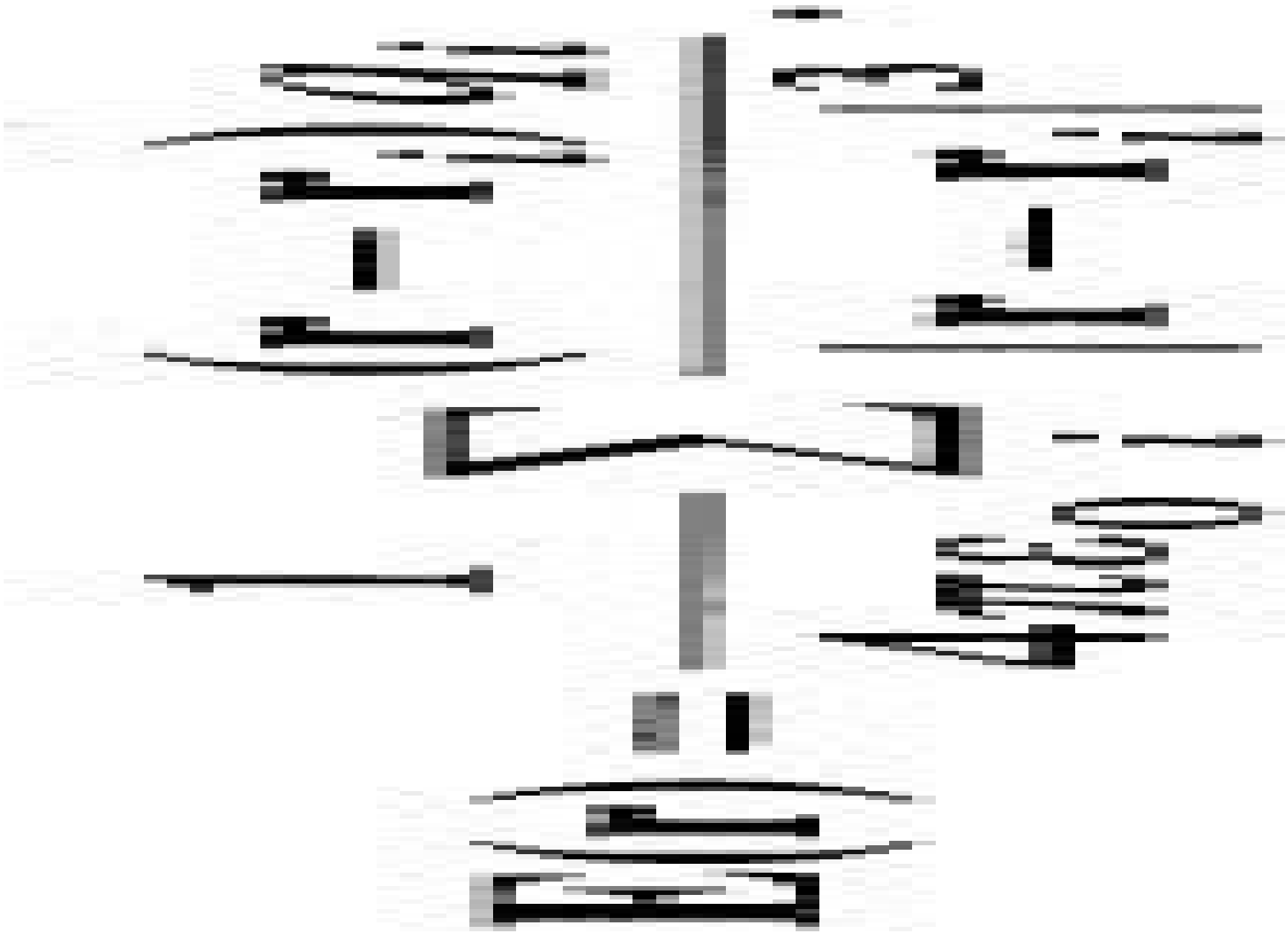
Here we have introduced a shorthand notation for the volume integral, which is really a triple integral over the three components needed to specify position vectors r within the volume V. In Cartesian coordinates $d\tau$ is dx dy dz, the volume of the rectangular box enclosed by sides of length dx dy dz. The volume of integration V must include the whole atom, but since $\rho_{el}(r)$ is zero outside the atom, it can be made as large as we please without affecting the result. If we extend V to cover the whole of space, in Cartesian coordinates the integral becomes



In atoms more complex than hydrogen, the electron clouds of different electrons overlap, and a volume element may contain parts of the charge of more than one electron. Even the charge carried by the nucleus should be represented by a charge density since although the nucleus is small compared with the whole atom, it does have a definite size. Now we define the atomic charge density $\rho_{atomic}(r)$ in a piece of matter as the net charge density at r, including positive contributions from nuclei. Where the charge densities associated with different particles overlap, they are added together to form the atomic charge density. The atomic charge density has large positive values inside the nucleus of each atom and is negative in the electron cloud. The volume integral $\int_V \rho_{atomic}(r)d\tau$ represents the net charge within V: thus for a volume V containing an electrically neutral piece of matter, the value of the integral must be zero.

1.3.2 The atomic electric field

The electric field due to an assembly of point charges qi was found by making a vector sum of the fields due to each charge separately:



Applying the same procedure to a continuous charge density ρatomic, the sum becomes an integral, and the atomic electric field at a point P with position vector r is

(1.7)

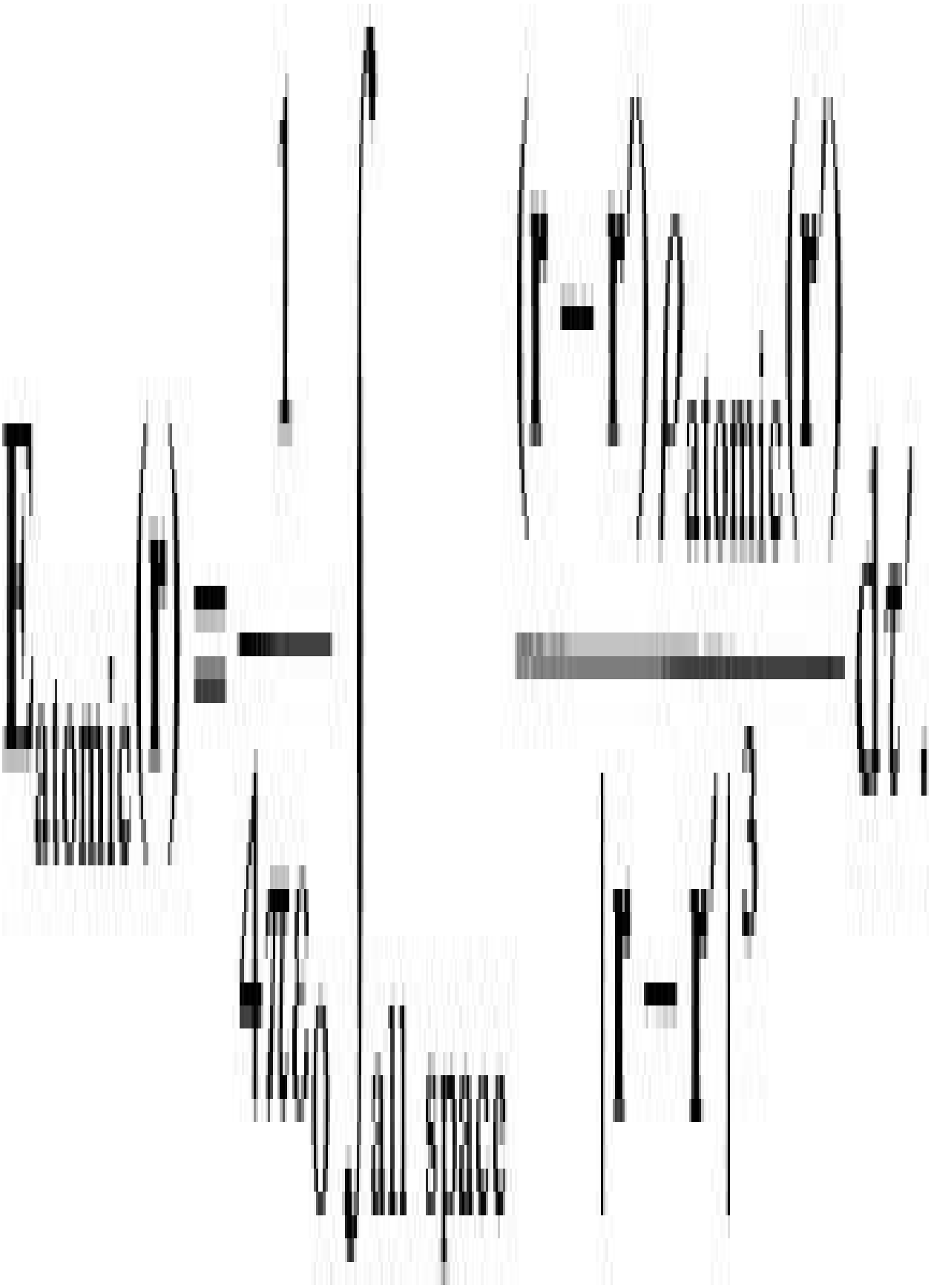
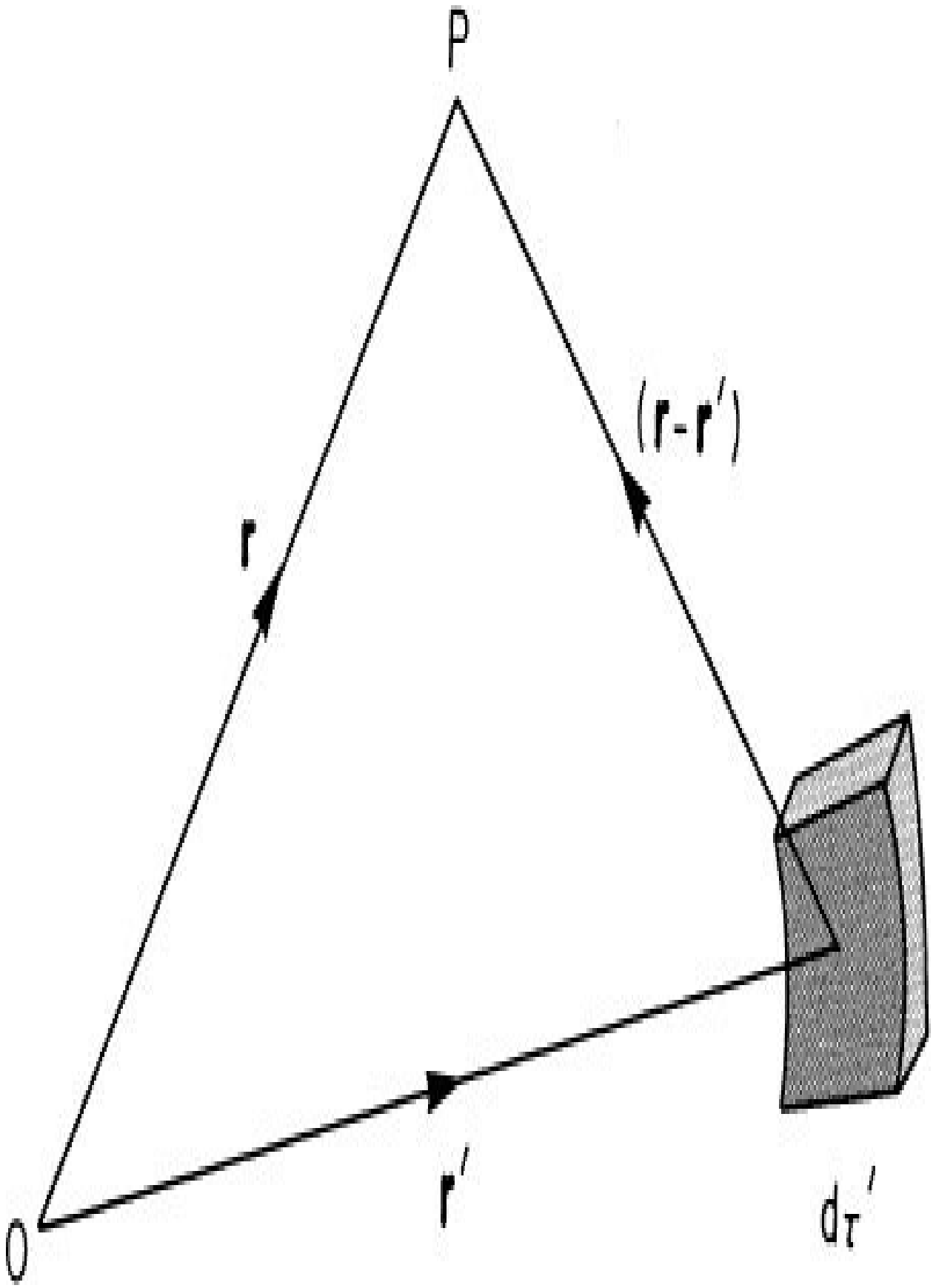


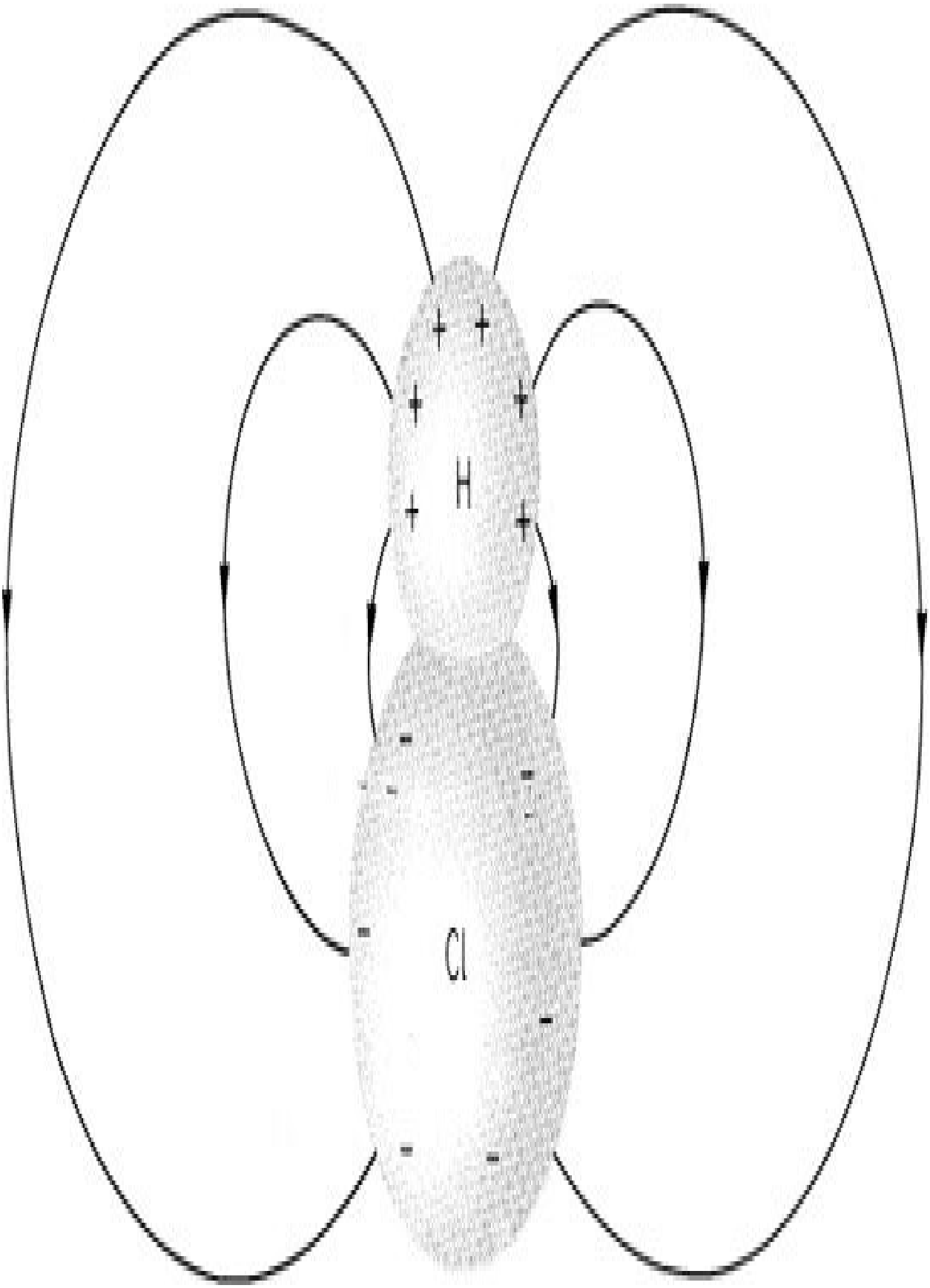
Figure 1.6. The charges within each volume element $d\tau'$ contributes to the electric field at P.


P
(r-r′)
r
r′
O
dτ′

Here ρatomic(r′) dτ′ is the charge within a volume element dτ′ located at the point with position vector r′, as illustrated in Figure 1.6. Both sides of Equation (1.7) are vectors, and to work out the volume integral in practice it is necessary to integrate each component separately.

Although it is easy to write down this equation for the atomic field*, it is not very useful, because in practice one does not know the charge density ρatomic. Not only does the charge density within each atom have a complicated form, but even in a solid, atoms are always in thermal motion. The charge density ρatomic and the field Eatomic at a fixed point are therefore continually changing. However, although the exact atomic field is rarely known, it is possible to describe its main qualitative features. Each atom has an internal field generated by its own nucleus and electron cloud. This field is very strong near the concentration of positive charge in the atomic nucleus. Superimposed on the internal field there may be fields caused by other atoms and molecules. Not only do charged atoms and molecules generate external electric fields, but even outside neutral atoms and molecules there may be fields which are significant over distances several times as large as a typical atomic diameter. A hydrogen chloride molecule, for example, is electrically neutral, but in the chemical bond which holds it together part of the electron cloud of the hydrogen atom is transferred to the chlorine atom. The hydrogen atom is thus left with an excess of positive charge, while the chlorine acquires a negative charge. Lines of force leave the positively charged hydrogen and end on the negatively charged chlorine as sketched in Figure 1.7 (notice that the lines of force indicate that the electric field outside the molecule has a similar shape to the field at some distance from the dipole in Figure 1.5). Fields of this kind fall off with distance much more quickly than the fields outside charged atoms and molecules. Nevertheless the local average field acting on a particular atom is often dominated by the contributions from neutral neighbours.

Figure 1.7. The field around a hydrogen chloride molecule: the hydrogen atom has lost part of its electron cloud and carries a net positive charge.


H
Cl

1.3.3 The macroscopic electric field

Fortunately, when one is considering large systems, the details of the atomic field are not important, and all that is needed is an average field. The value of the field averaged over a region much bigger than a single atom, but still small compared to the size of the whole system, is called the macroscopic electric field (the word macroscopic means large-scale, and is used to make a contrast with the microscopic scale of the atomic field). In the macroscopic field all the sharp variations of the atomic field within each atom are smoothed out: the macroscopic field is the one which would exist if matter were continuous and had no atomic structure.

By comparison with the strong atomic fields near nuclei, macroscopic fields are of modest magnitude. In a good electrical insulator like porcelain, electrons are bound tightly to their atoms, and any local excesses of positive or negative charge can remain immobile for a long time, maintaining a macroscopic field. But if this field becomes too large catastrophic breakdown occurs. The largest fields which occur are about 10^9 volts/metre. Static fields as great as this do not occur over extensive regions, and even aver distances of a few atomic diameters, they are found only in some special situations. *.

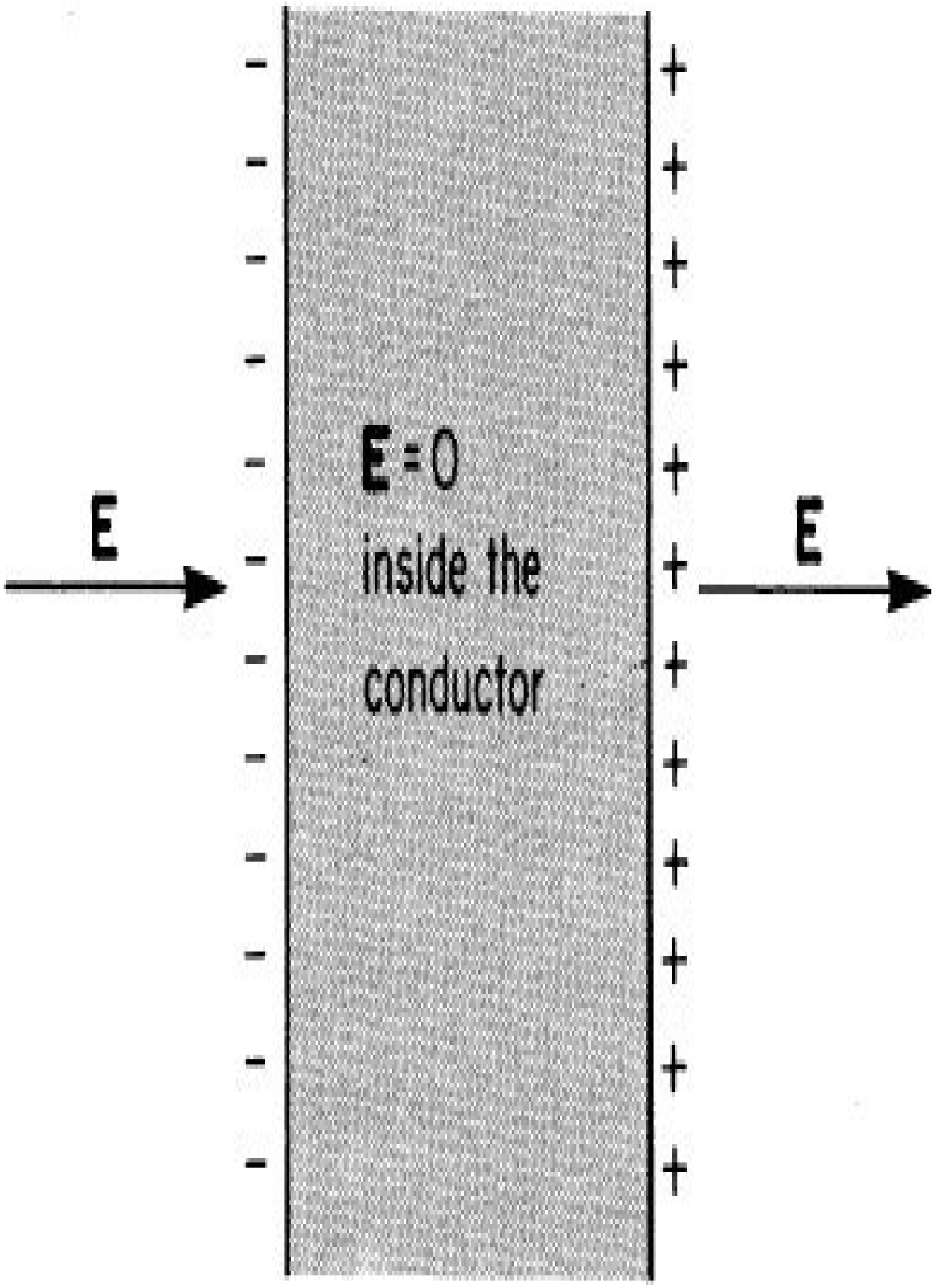
Net charges may be distributed throughout the volume of an insulator, or may be concentrated in a thin layer at a surface. Only surface charges occur in a material like copper which is a good electrical conductor. In a conductor, some electrons, which are referred to as conduction electrons, are free to move in the presence of a macroscopic electric field. It follows that in static equilibrium the macroscopic electric field within a conductor is everywhere zero. What happens when a conductor is placed in an electric field generated by outside charges? Consider a slab of conductor in a uniform field directed perpendicularly to the surface of the slab. When the field is first applied, a macroscopic field extends throughout the conductor. The conduction electrons move under the influence of this field,

leaving net charges on the surface as shown in Figure 1.8. The surface charges generate a field inside the conductor which is opposed to the external field, and the surface charges build up until the macroscopic field within the bulk of the conductor is exactly zero. Charges appearing on the surface of a material in this way, due to the presence of an external field, are called induced charges.

We now relate the macroscopic field to the charges which generate it, taking into account both distributed volume charges and surface charges. First consider the volume charges. Imagine a volume δV, small compared with a macroscopic system, but still large enough to contain many atoms. The net charge within δV is $\int_{\delta V} \rho_{atomic}(r) d\tau$. Neighbouring regions of the same volume δV may not contain precisely the same net amount of charge. However, the relative differences between regions are small, provided that each region contains a large enough number of atoms. We can define the macroscopic charge density $\rho(r)$ at the point with position vector r as the net charge per unit volume contained in a volume δV in the vicinity of r, i.e.

(1.8)

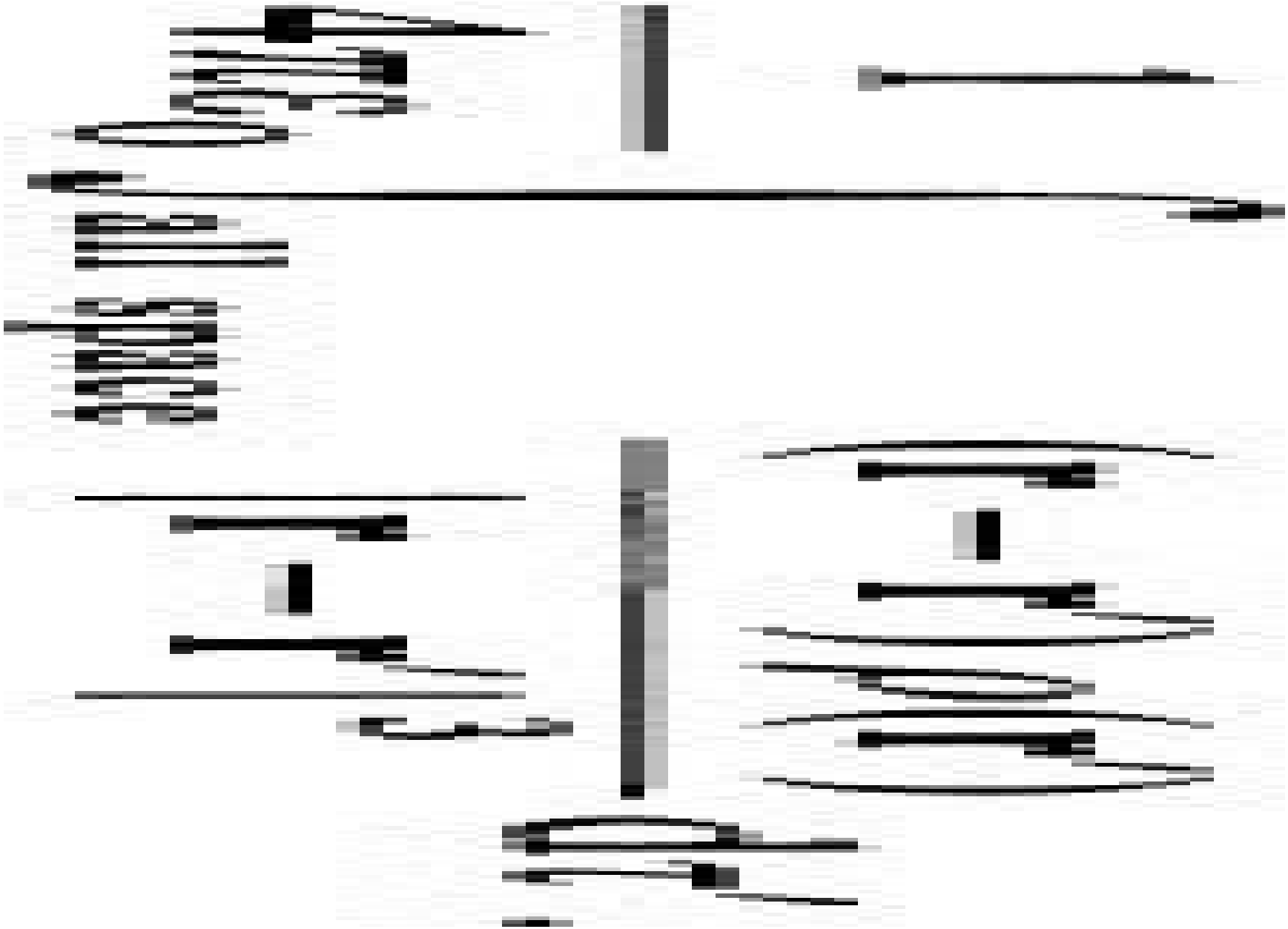
Figure 1.8. Induced surface charges appear on a slab of conductor when it is placed in an electric field.


E
E = 0
inside the
conductor
E

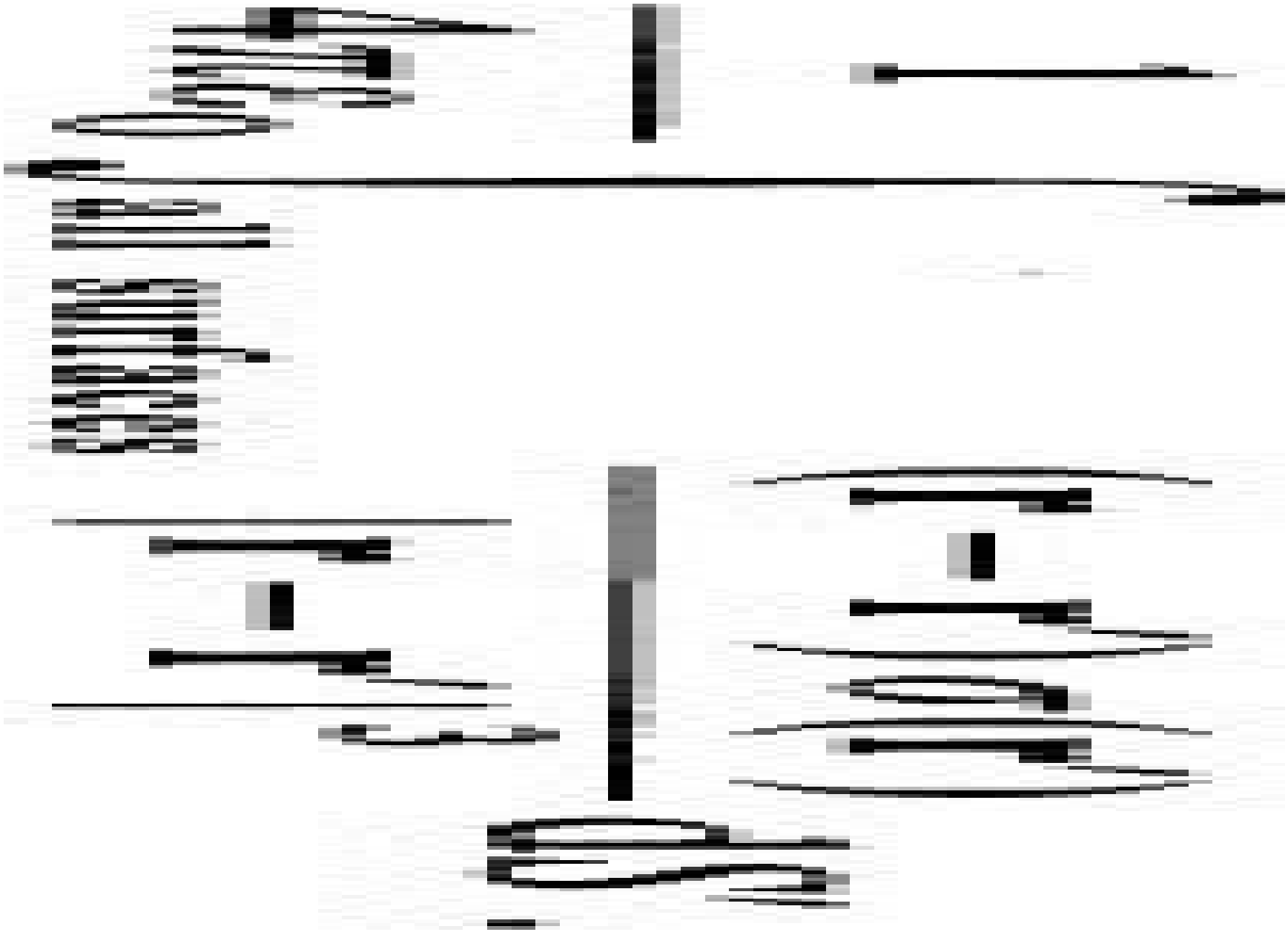
The charge density ρ(r) is a function of position which varies smoothly inside matter, except at boundaries between media, where ρ may change discontinuously. Near a boundary there may also be a concentration of surface charges. Surface charges always occupy a very thin layer, no more than a few atoms thick, and for the purpose of discussing the macroscopic field they may be represented by an infinitesimally thin sheet of charge lying exactly on the boundary surface. A surface charge density σ(r) is defined, analogously to the volume density ρ(r), as the net charge per unit area on an area δS which is small on the macroscopic scale, but large enough to include many surface atoms. Consider a slab of volume δV just thick enough to enclose all the surface charge σ(r)δS associated with an area δS in the vicinity of a point on the surface with position vector r. The net charge within δV is a volume integral over the atomic charge density, and we have*

(1.9)

Macroscopic electric fields obey the principle of superposition just as do atomic fields. To find the macroscopic field at a point with position vector r, it is necessary to add the contributions from the net charges ρ(r′) dτ′ in all the volume elements dτ′ covering the whole of space, and the contributions from the surface charges σ(r)dS′ on the surface elements dS′ over all boundary surfaces. The field generated by the volume charge density ρ is given by a volume integral of the same form as Equation (1.7) for the atomic field, namely



Similarly the surface charge density σ generates a field given in shorthand notation as



The surface integral is a double integral, and for example if the rectangular surface in the x–y plane shown in Figure 1.9 carries a charge density σ (and has corners at $x = \pm a$, $y = \pm b$), its contribution to the field at a point with position vector r is

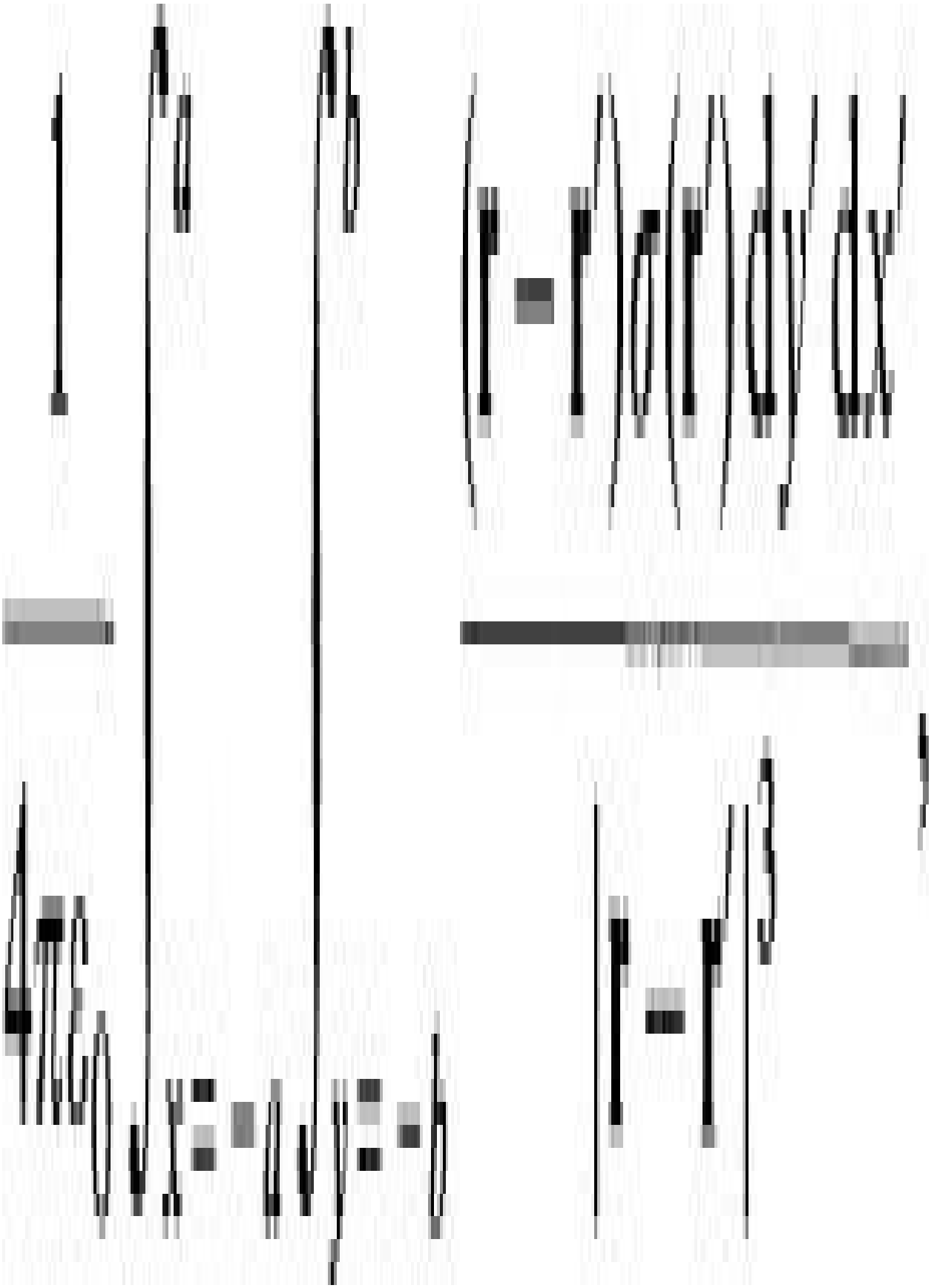
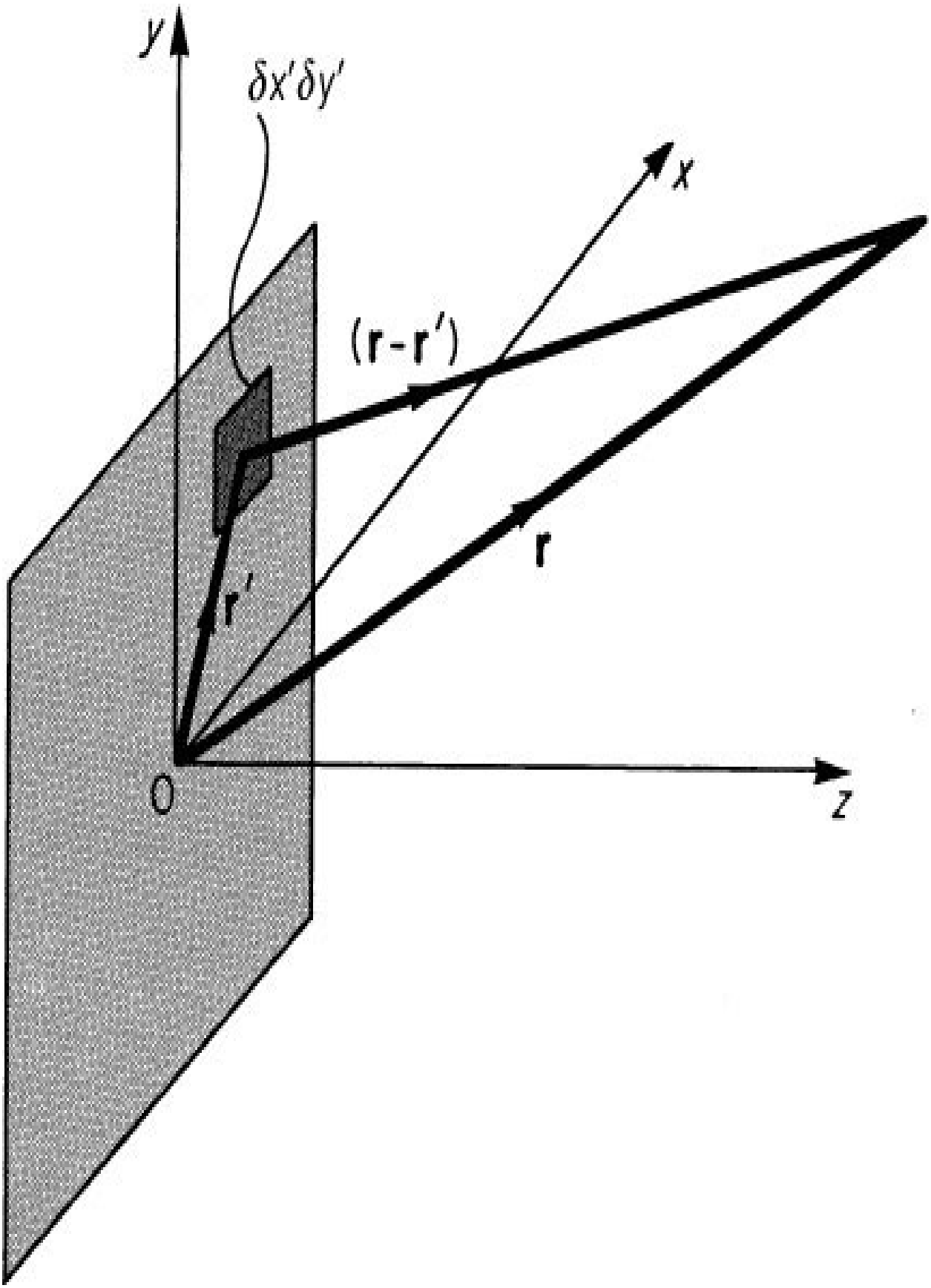
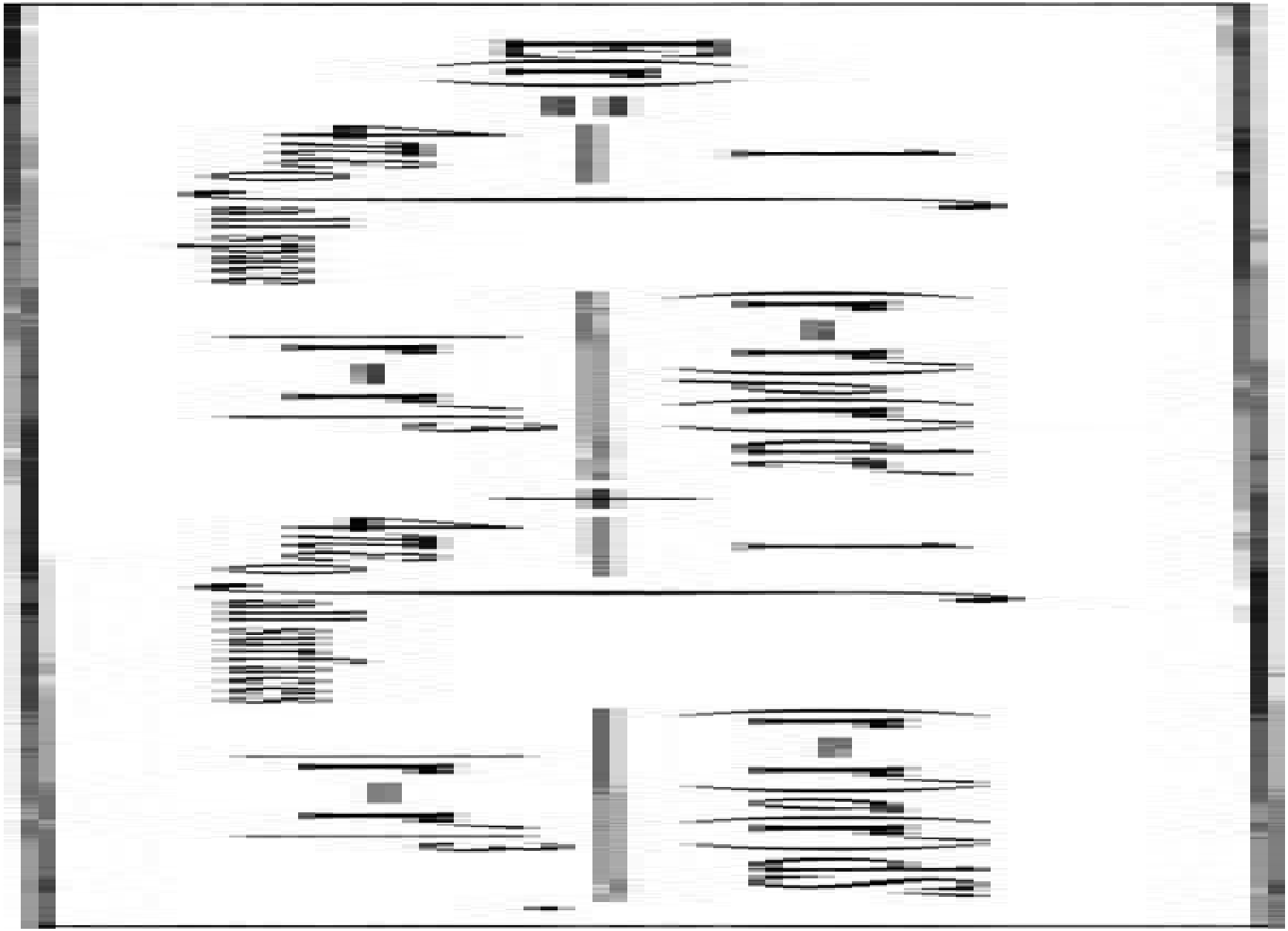


Figure 1.9. The vector addition of the fields due to all surface elements $\delta x' \, \delta y'$ leads to a surface integral.


y
δx′δy′
x
(r-r′)
r
r′
0
z

Finally, adding the terms from volume and surface charge densities, the macroscopic electric field is defined to be

(1.10)



1.4 GAUSS' LAW

When the charge density has a distribution with a simple symmetry, the electric field can be evaluated easily by the application of Gauss' law, an important result which relates the field on any closed surface to the net amount of charge enclosed within the surface. Using Gauss' law we shall be able to prove, for example, that the field outside an isolated and spherically symmetrical charged ion is exactly the same as if all its charge were concentrated at the centre; an application to a macroscopic system is at the surface of a conductor, where Gauss' law leads to an expression for the external electric field in terms of the surface charge density. Gauss' law can be illustrated by considering a point charge q at the centre of a sphere of radius r. The area of the sphere is $4\pi r^2$, and the electric field on its surface is everywhere of magnitude $q/4\pi\varepsilon_0 r^2$. The product of electric field and area is q/ε_0, i.e. it is independent of r. This is rather like what happens when a light bulb is placed at the centre of a spherical lampshade which is completely transparent. All the light from the bulb escapes, or in other words the total amount of light passing through the shade is independent of its radius. Obviously the total amount of light escaping is also independent of the shape of the shade; all the light gets out whether or not the shade is spherical. Gauss' law is the analogous result in electrostatics applied to a quantity called the flux of the electric field; the total flux out of a surface enclosing a charge q is q/ε_0, whatever the shape of the surface. In the rest of this section we shall explain what is meant by flux, and then go on to prove Gauss' law rigorously.

1.4.1 The flux of a vector field

Flux may be defined for any vector function of position, but it is most easily visualized in the flow of fluids. Imagine that a fluid is flowing at a speed V and that a small loop of wire is placed in the fluid. In Figure 1.10(a) the shaded area δS, which is bounded by the loop of wire, is perpendicular to the direction of flow of the fluid. The flux of fluid through the area δS is the rate of flow of fluid through the loop, which in this case is the product $v\,\delta S$ of speed and area. Now

suppose that the loop is rotated to the position shown in Figure 1.10(b), where the normal to the surface δS is at an angle Ψ to the direction of flow. Looking along the direction of flow, the projected area within the loop is now only $\delta S_p = \delta S \cos \Psi$, and the flux, i.e. the rate of flow of fluid through the loop, is reduced to

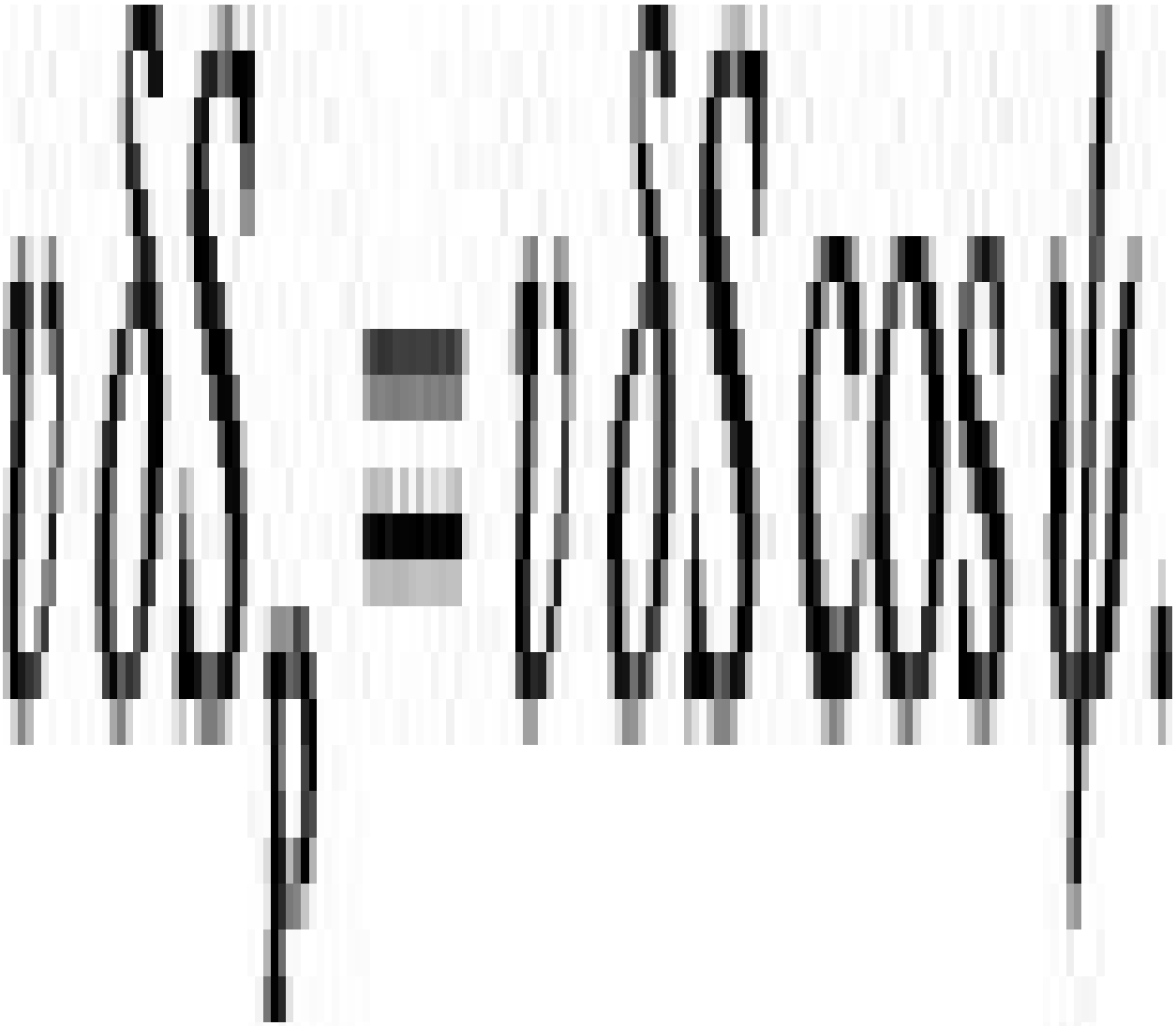
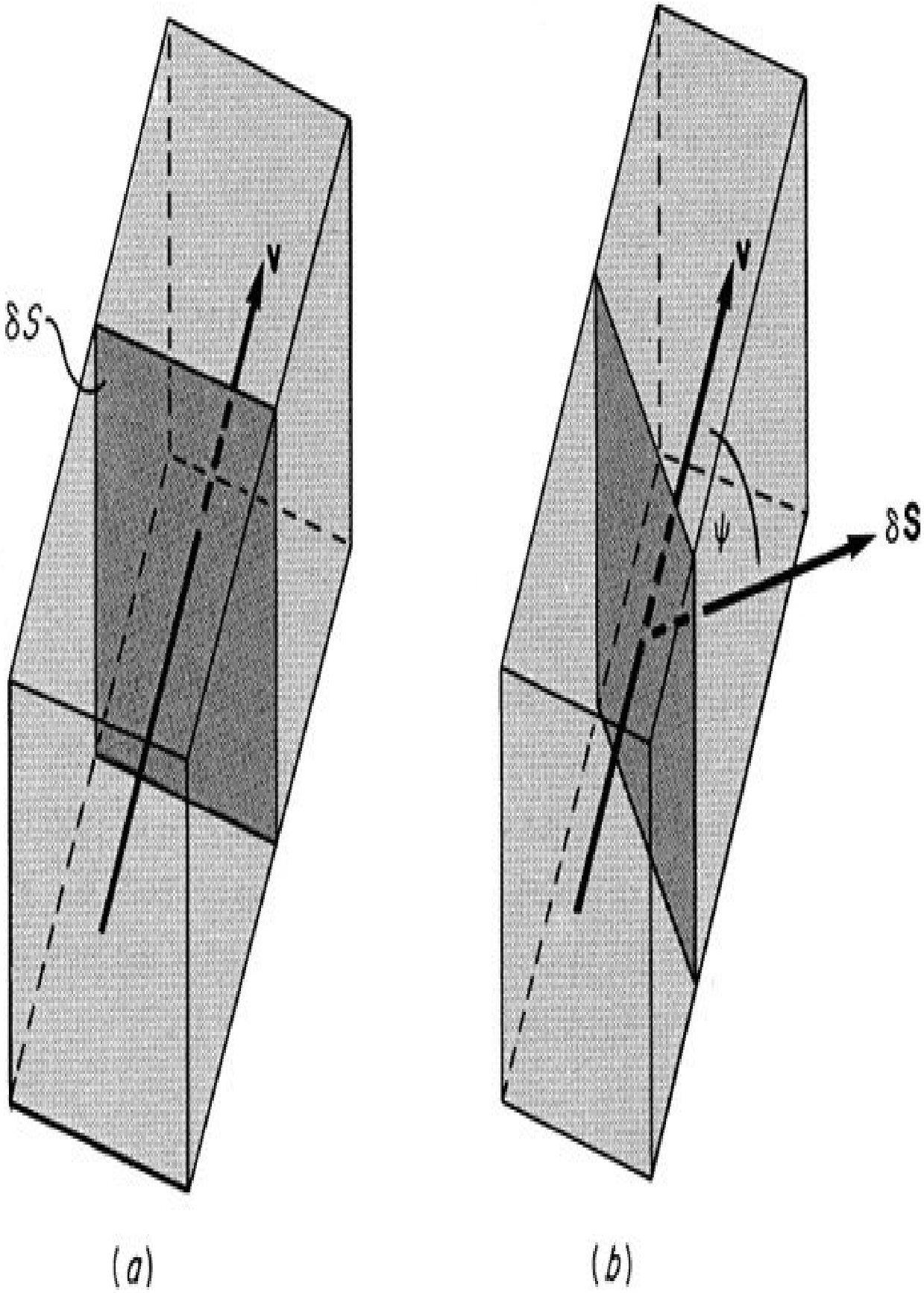
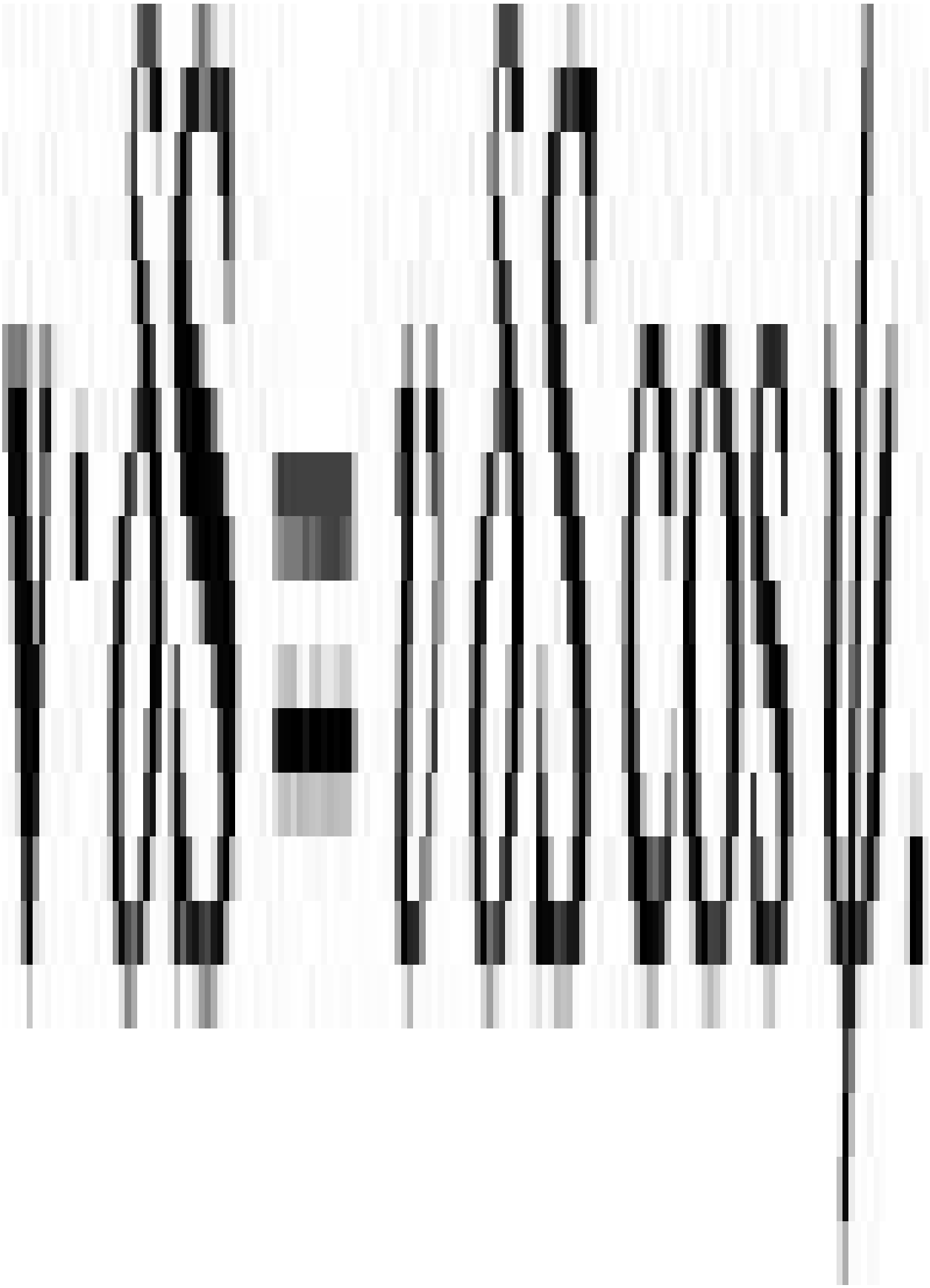


Figure 1.10. A column of fluid of length v passes through the area δS every second. The volume of fluid passing is larger when v is perpendicular to δS, as in (a), than when it is at an angle as in (b).


δS
V
(a)
V
ψ
δS
(b)

This flux can be written in vector notation if we represent the surface by a vector δS, of magnitude δS and directed along the normal to the surface. The velocity vector v is then at an angle Ψ to δS, and the flux through δS is given by the scalar product



The flux has a sense as well as a magnitude, and in Figure 1.10(b) the sense is from the left-hand side of the loop to the right-hand side. If we choose to represent the surface by the vector $(-\delta S)$, still normal to the surface but pointing from right to left in the opposite direction to δS, the scalar product $v\cdot(-\delta S)$ becomes negative. A negative flux from right to left is equivalent to a positive flux from left to right, but to avoid ambiguity the sense of surface vectors must always be carefully specified.

In an electrostatic field there is of course nothing actually flowing, but the flux of an electric field E through δS is defined in the same way as $E\cdot\delta S$, the value of E being taken at the position of the surface δS. In terms of components the flux of E through δS can be expressed as

$E\cdot\delta S$ = (magnitude of E) × (projected area δS_p)

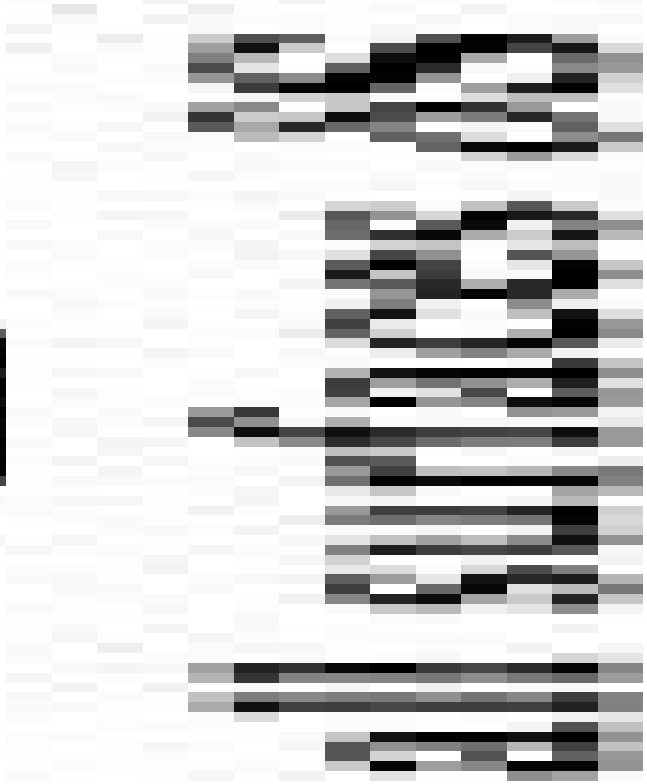
= (magnitude of E) × (component of δS along the direction of E),

or equivalently

$E\cdot\delta S$ = (component of E normal to the surface) × (area δS).

So far we have only defined the flux through a small flat surface. To find the flux of the electric field through a surface S of arbitrary shape, we first approximate the shape of S by dividing it up into a lot of small flat surfaces. The flat surfaces are represented by vectors such as δS in Figure 1.11, all pointing outwards from the same side of the whole surface. The flux of the field E through δS is $E\cdot\delta S$,

and the total flux through the surface made up of the mosaic of flat surfaces is the sum of the fluxes through each separately, i.e.



If the size of each of the flat surfaces is made smaller and smaller, the mosaic approaches more and more closely to the smooth surface S, and the sumapproaches a limiting value which is equal to the flux through S. The limit is a two-dimensional surface integral written as

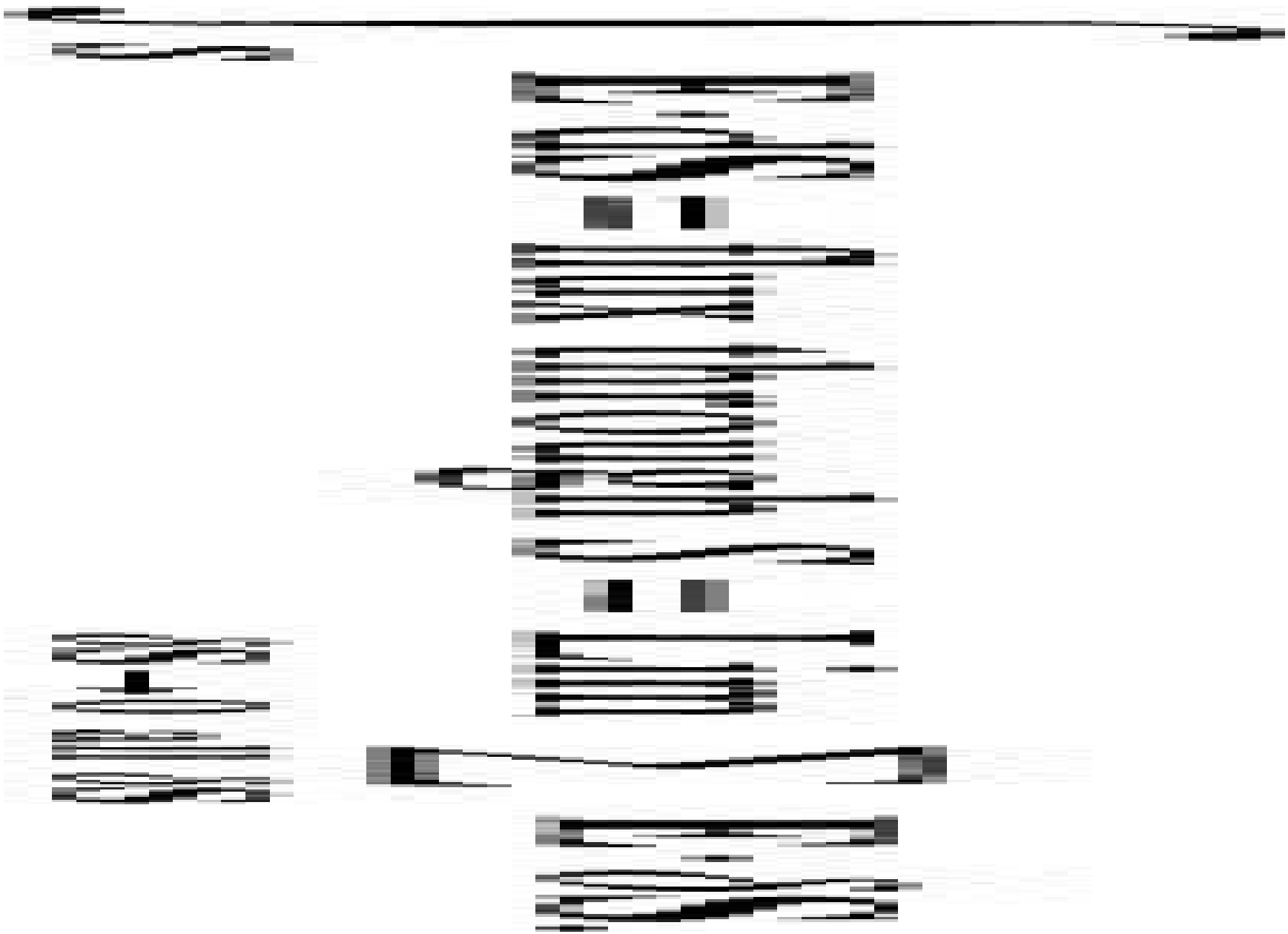
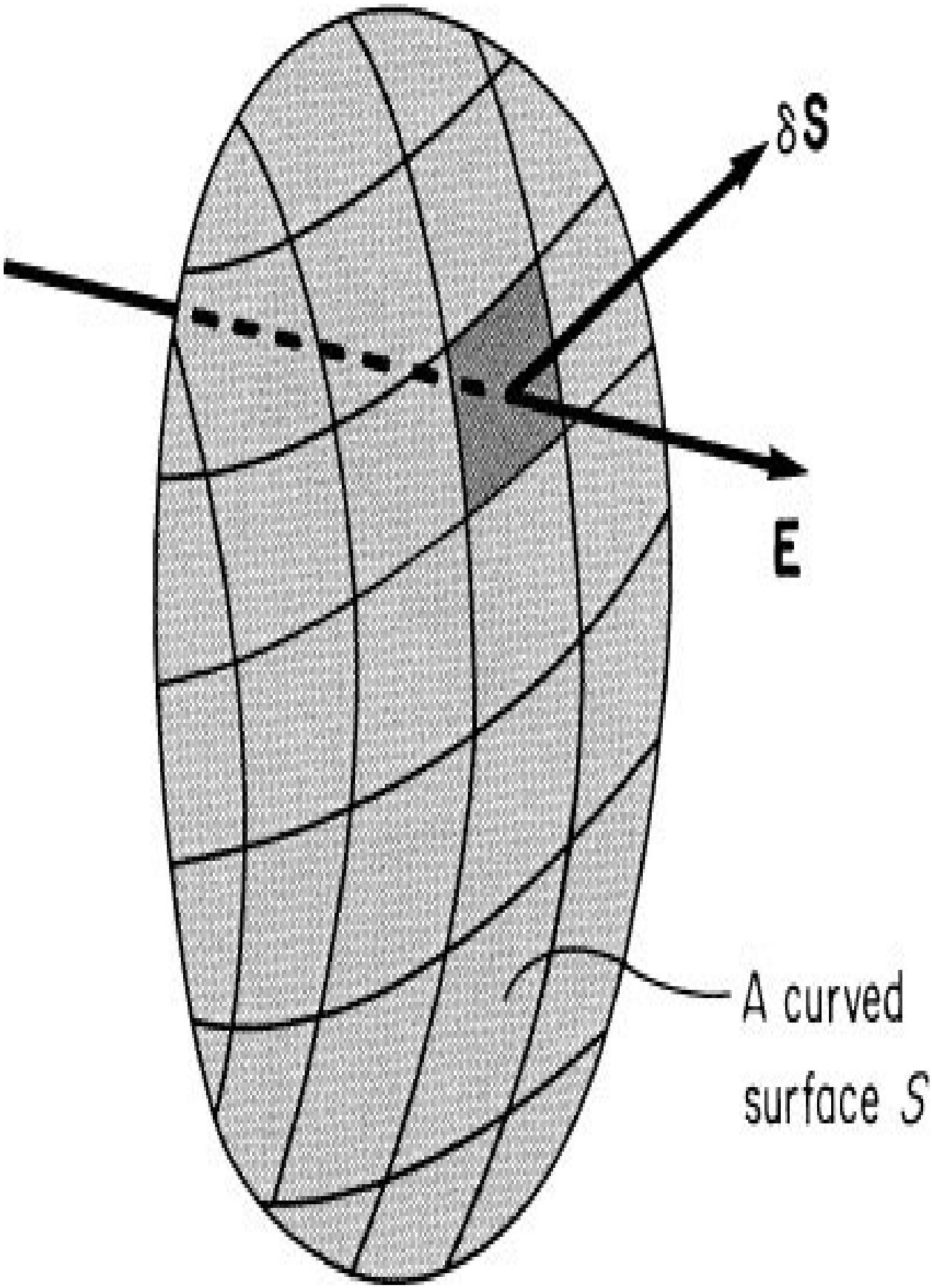


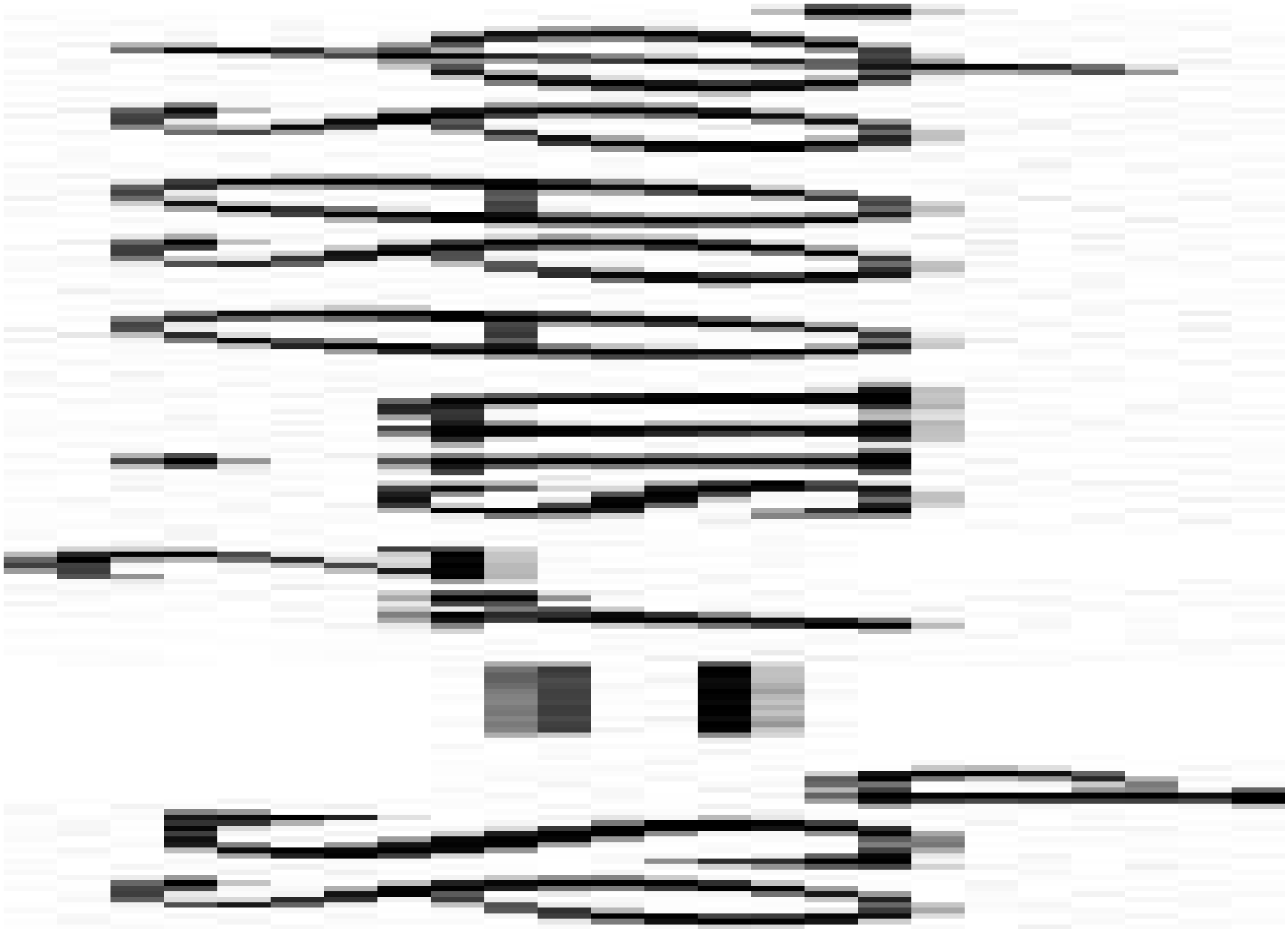
Figure 1.11. Splitting up a surface S into a lot of little flat surfaces in order to work out the flux of E through S.


δS
E
A curved
surface S

The subscript S under the integral sign indicates that the area of integration is the whole of the surface S.

1.4.2 The flux of the electric field out of a closed surface

Let us now evaluate the flux through a closed surface S for the electric field E generated by a point charge q enclosed by S. We shall use spherical polar coordinates referred to an origin at q, and concentrate attention for the moment on a surface element of area δS at the position with coordinates (r, θ, Φ). The surface element is represented by the vector δS directed along the outward normal to S. The element is of such a size and shape that its projection along the radius vector from q lies between the polar angles θ and $(\theta + \delta\theta)$ and the azimuthal angles Φ and $(\Phi + \delta\Phi)$. As illustrated in Figure 1.12, the projected area is



According to Equation (1.5), the field at δS is directed outwards along the radius vector and has a magnitude

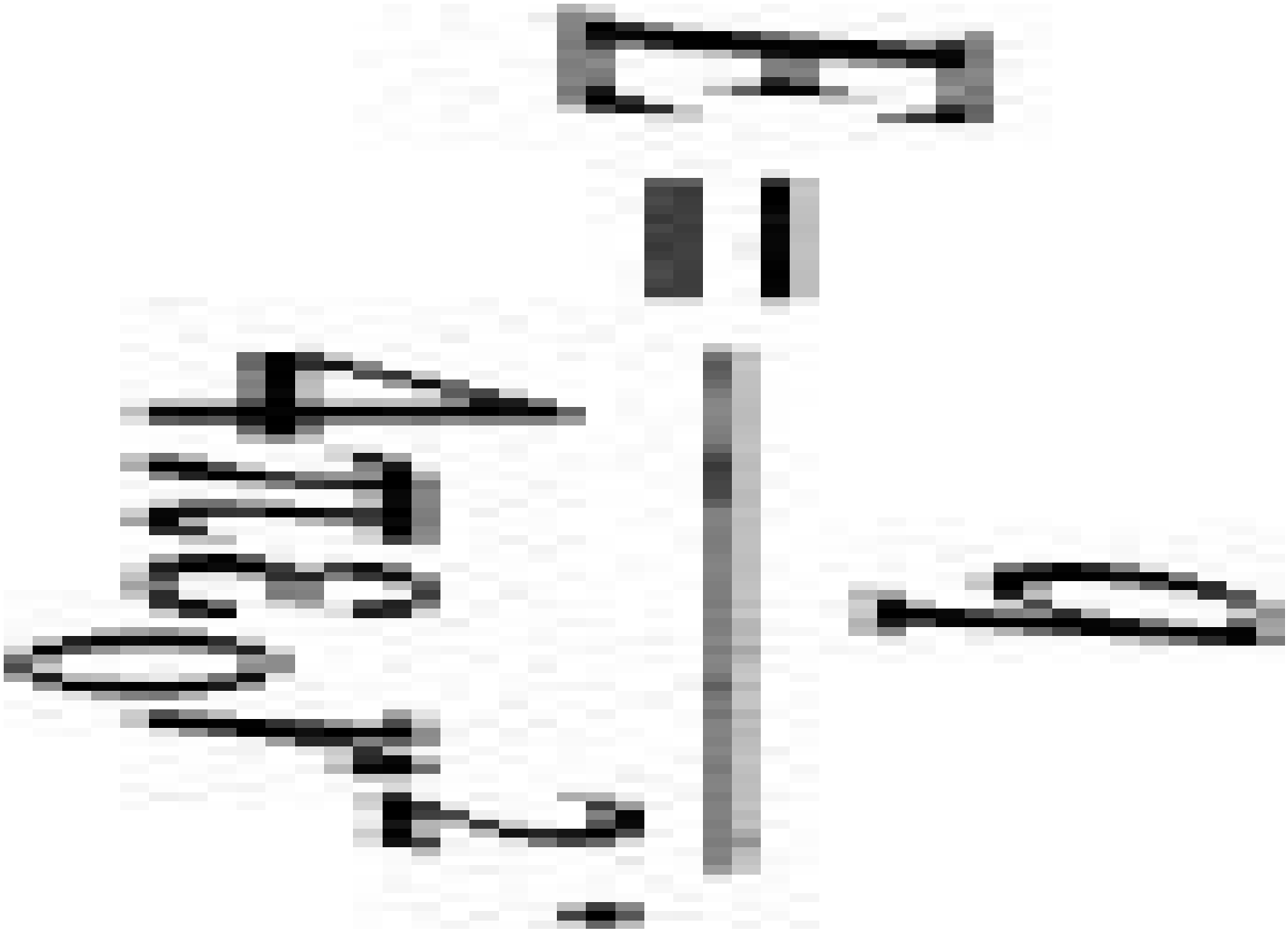
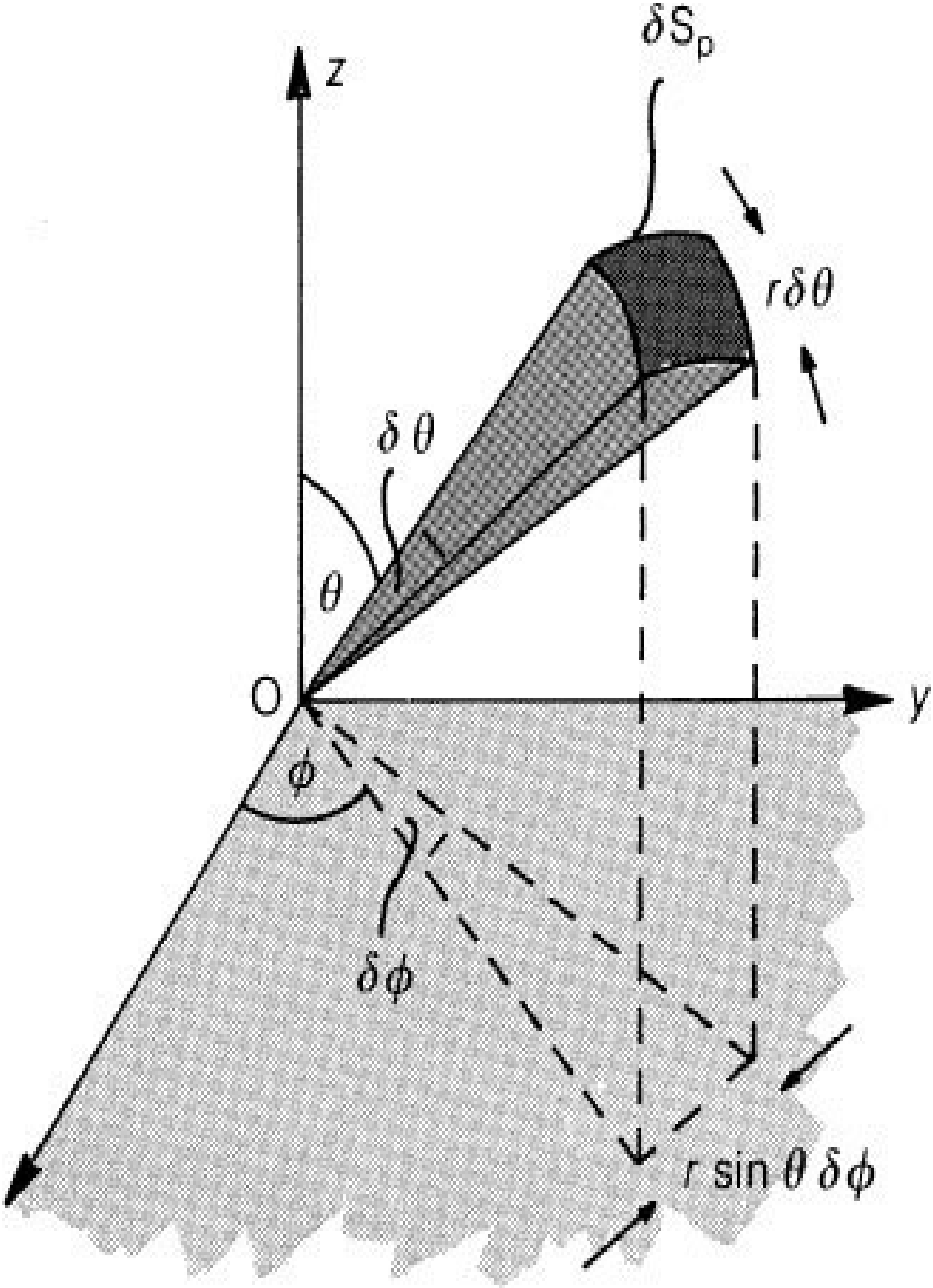


Figure 1.12. The area δSp subtends angles $\delta\theta$ and $\delta\varphi$ at the origin. It is a distance r from the origin, so that it is a rectangle of sides $r\delta\theta$ and $r \sin\theta\delta\varphi$, i.e. $\delta SP = r^2 \sin \delta\theta \, \delta\varphi$.


δS_p
z
$r\delta\theta$
$\delta\theta$
θ
O
y
ϕ
$\delta\phi$
$r \sin\theta\, \delta\phi$

Hence the outward flux through δS is

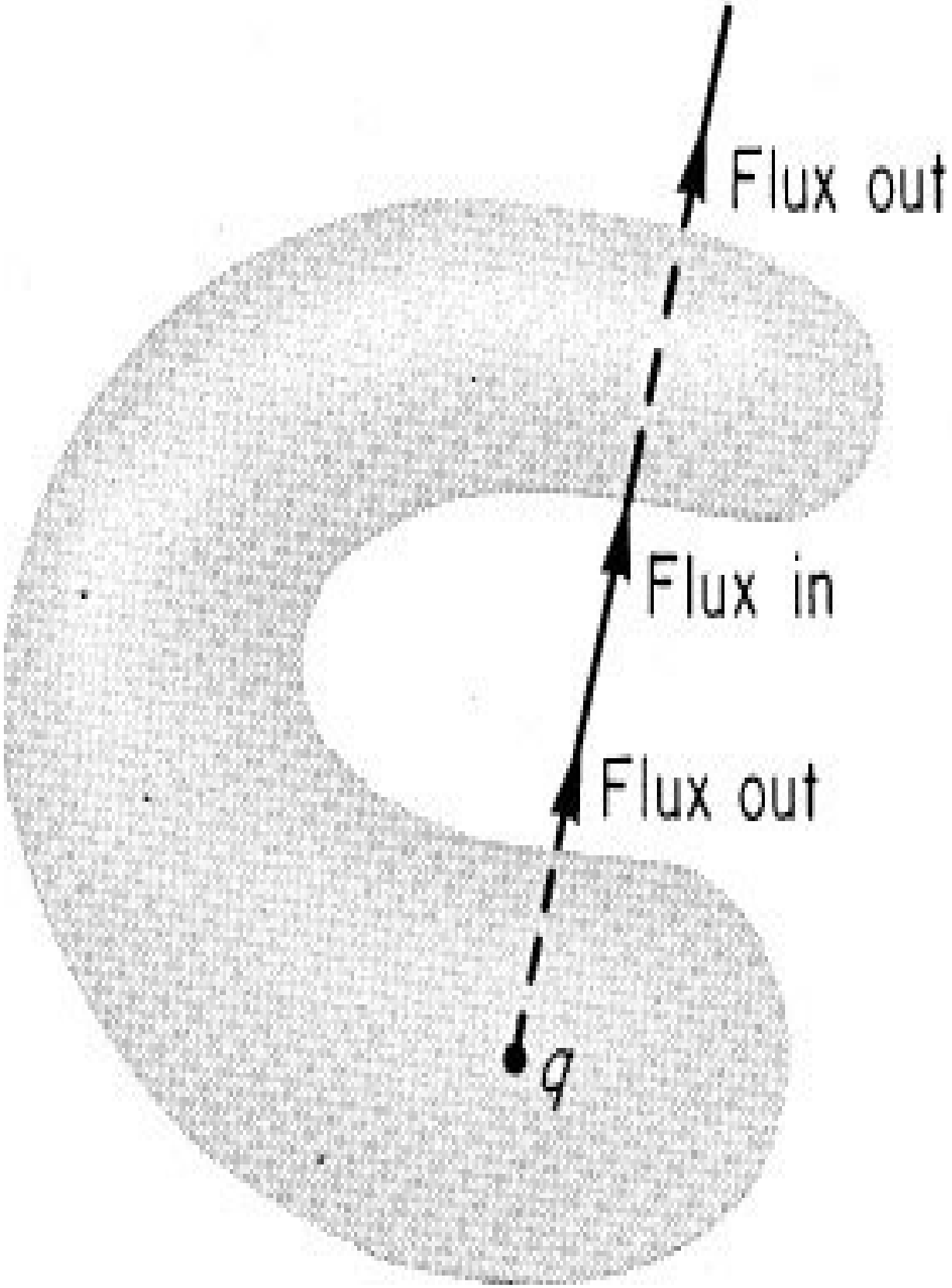
(1.11)

where the element δΩ. is termed the solid angle subtended by δSp at the origin. Using Equation (1.11), we can now transform the surface integral expression for the flux through the complete surface S into an integral over angular variables only. The range of variation required to cover the whole of the closed surface is θ = 0 to π and φ = 0 to 2π. In the limit, as δS tends to zero, the total flux through S is

(1.12)

In this derivation it has been assumed that the flux through dS is always positive. Sometimes, however, when S has a complicated shape, the radius vector from q may pass from outside to inside the enclosing surface, as in Figure 1.13. Where the field is directed inwards it makes a negative contribution to the net outward flux. Provided that the direction of dS is chosen to be along the outward normal to S, the scalar product E·dS always gives the correct sign for the flux out of dS. But Equation (1.11) should be modified to

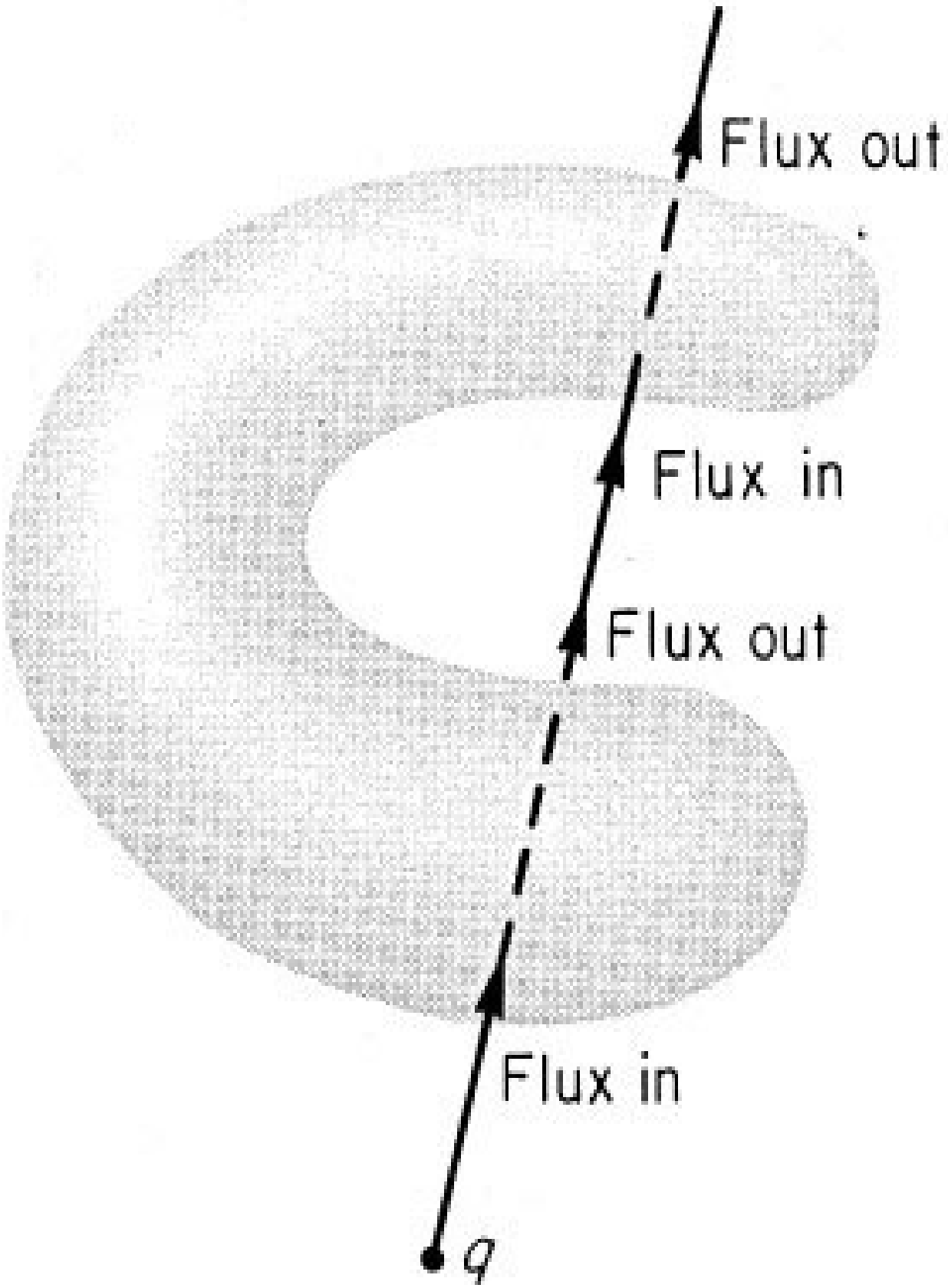
Figure 1.13. The radius vector from a charge inside a surface always makes one more outward crossing than the number of inward crossings.


Flux out
Flux in
Flux out
q

(1.13)

with the negative sign applying for those parts of the surface where the field is directed inwards. Notice that however many times a radius vector may cross the surface S, if the origin lies within S the radius vector must always make one more outward crossing than the number of inward crossings. It follows that no matter how many times the cone of solid angle dΩ, is cut by the surface S, the net flux leaving S within dΩ is $(q/4\pi\varepsilon_0)\, d\Omega$, and Equation (1.12) is correct for any surface whatever which encloses q.

Figure 1.14. For a charge outside the surface the number of outward crossings is the same as the number of inward crossings.


Flux out
Flux in
Flux out
Flux in
q

For a charge q which is outside a closed surface S, on the other hand, the radius vector must cross S an even number of times, with equal contributions to the inward and outward fluxes, as illustrated in Figure 1.14. The net outward flux through a closed surface due to any charge not enclosed by the surface is therefore exactly zero. Now we can use the principle of superposition to deduce the outward flux through a closed surface S of the electric field E due to an arbitrary distribution of point charges qi. Only those charges enclosed by S make a contribution to the flux, and

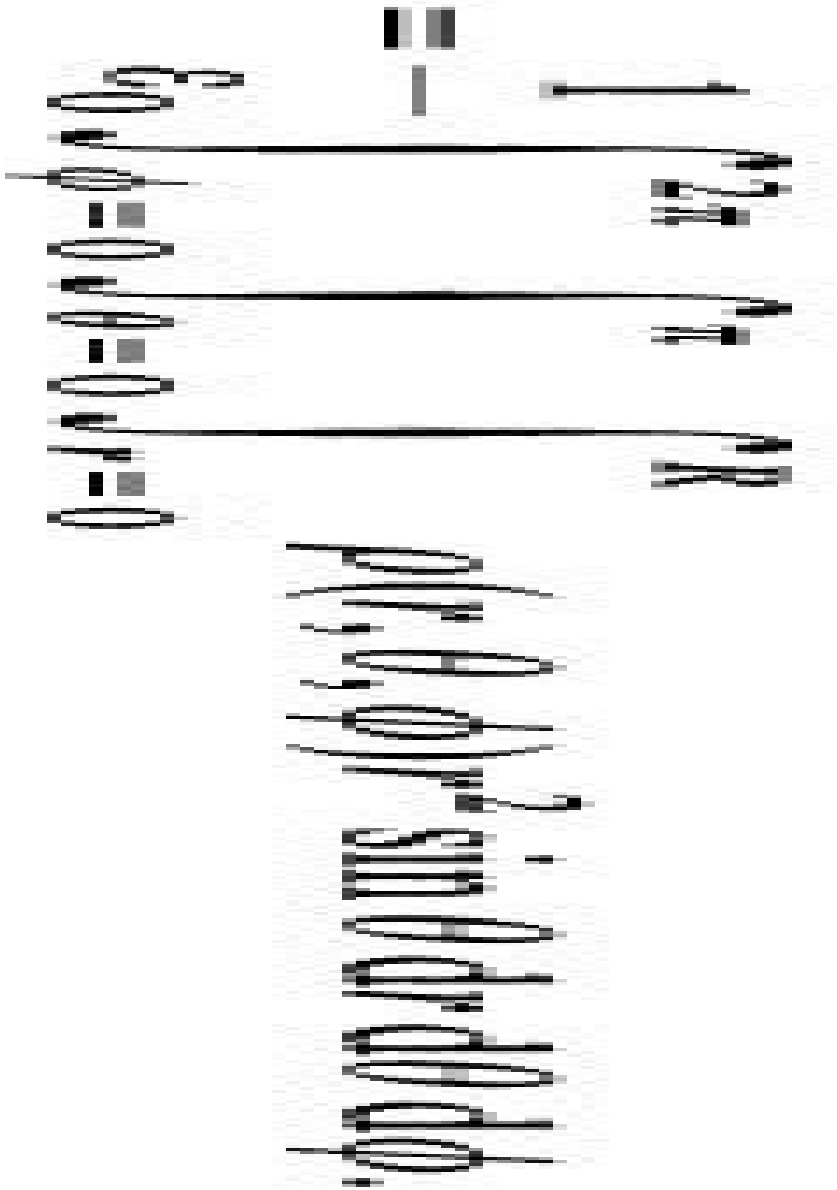
(1.14a)

where the summation is restricted to the charges within S. If the charge has a continuous distribution with a charge density ρ(r), then the summation is replaced by an integral, and

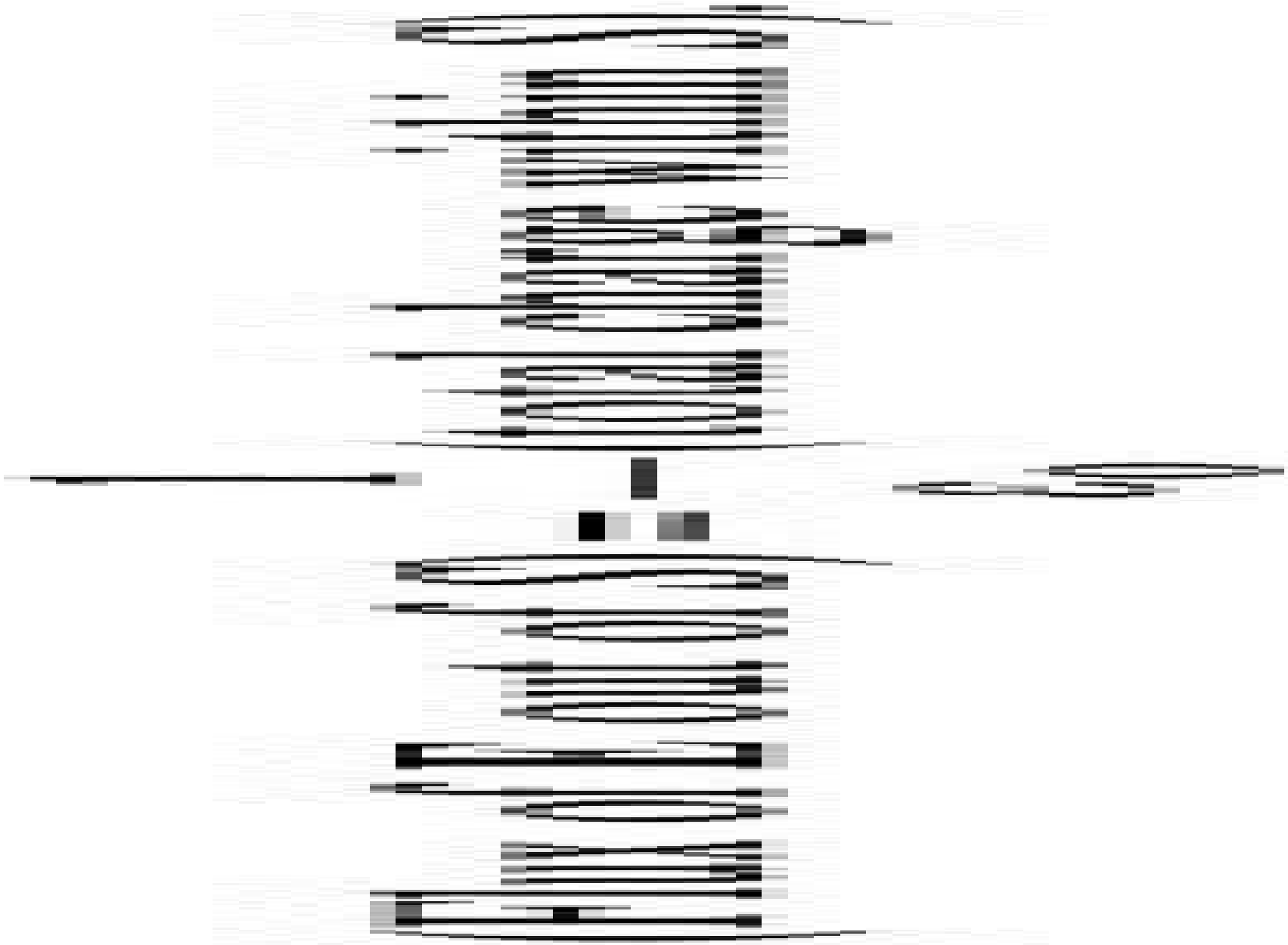
(1.14b)

where the integral on the right-hand side extends over the whole of the volume V enclosed by S. Here we have related a two-dimensional surface integral to a three-dimensional volume integral. To illustrate how to apply the relation to a particular coordinate system, let us write it out in full for a sphere of radius R. In spherical polar coordinates a position vector r is specified by the three variables r, θ and φ. The charge density and electric field are both functions of position, and are written as ρ(r, θ, φ) and E(r, θ, φ). The electric field at the surface of the sphere may not be normal to the surface, but only the radial component Er contributes to the flux of E through the sphere. The element of surface subtending angles dθ and dφ) at the origin is $R^2 \sin\theta \, d\theta \, d\varphi$ (see Figure 1.12), and the volume element within dr, dθ and dφ is $r^2 \sin\theta \, dr \, d\theta \, d\varphi$. With limits which extend the integrations over the whole of the surface and volume of the sphere Equation (1.14b) becomes

(1.15)

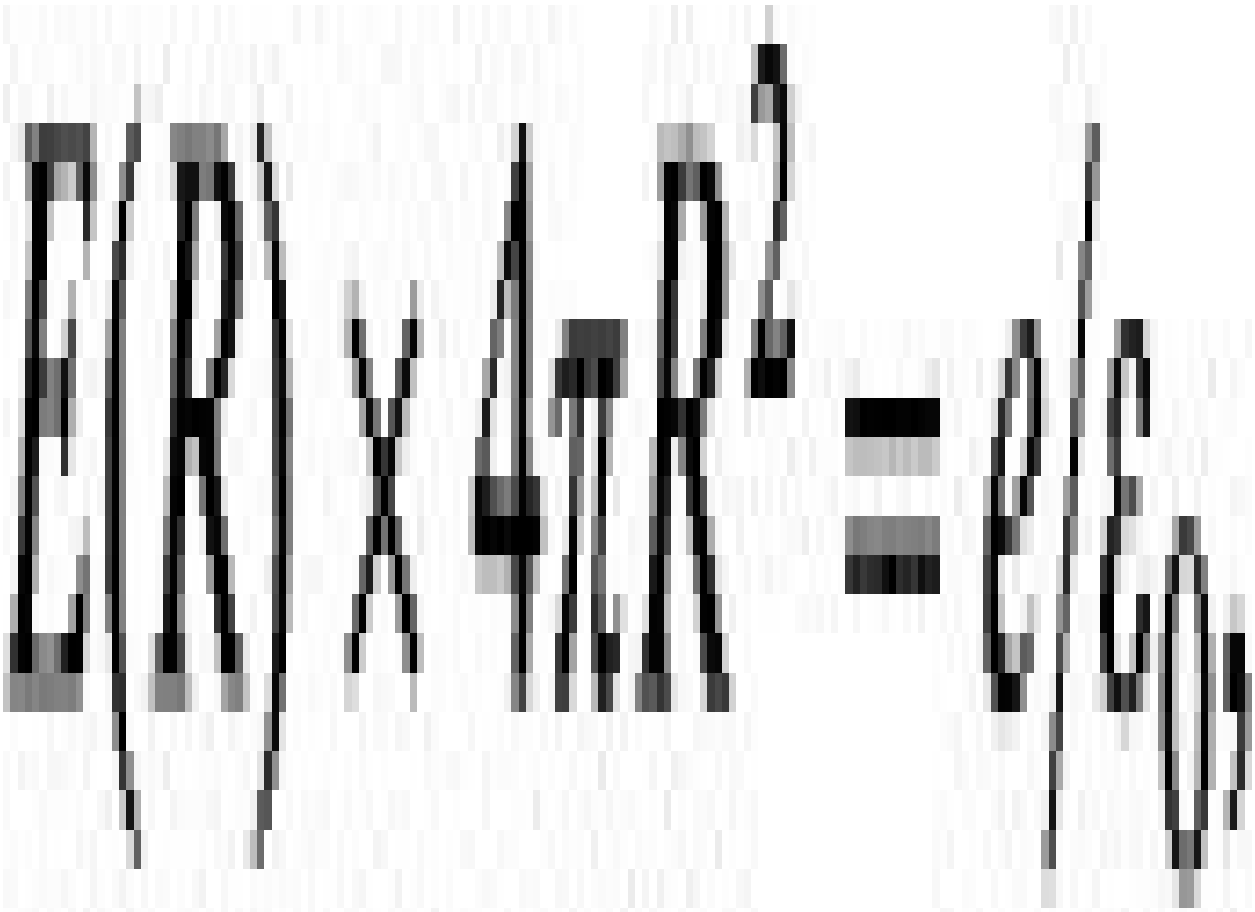


Whether it is applied to point charges or a continuous charge distribution, Equation (1.14) can be expressed in words as



This equation, which is Gauss' law, is a direct result of the inverse square dependence of the electric field. Only with an inverse square law does a point charge generate a flux per unit solid angle which is independent of distance from the source. The analogy between Gauss' law and the flux of light through a closed surface surrounding a lamp also rests on the inverse square law. Light intensity diminishes as the inverse square of the distance from the source of light, and in this case the constancy of flux confined within a fixed solid angle, as for example in the beam of a flashlight, is a consequence of the conservation of energy.

As has already been mentioned, Gauss' law can be used to calculate the field when the charge distribution has a simple symmetry. An example of spherical symmetry is provided by an isolated ion of singly ionized sodium, that is, a sodium atom which has lost one of its electrons and carries a net charge + e. The field outside the ion can be calculated at once using Gauss' law, without any knowledge of the distribution of charge inside. Choose a spherical surface of radius R centred on the ion, and large enough to contain the whole of the ion. Gauss' law then takes the form of Equation (1.15). Because of the spherical symmetry, the electric field on the surface does not depend on θ and φ, but has a constant magnitude E(R), say, and is directed radially away from the origin. Taking E(R) outside the surface integral, the integral over θ and φ is simply the area $4\pi R^2$ of the sphere, and the total flux is $E(R) \times 4\pi R^2$. The total charge within R must be + e, whatever may be the radial distribution of the charge carried by the electrons. Hence



or

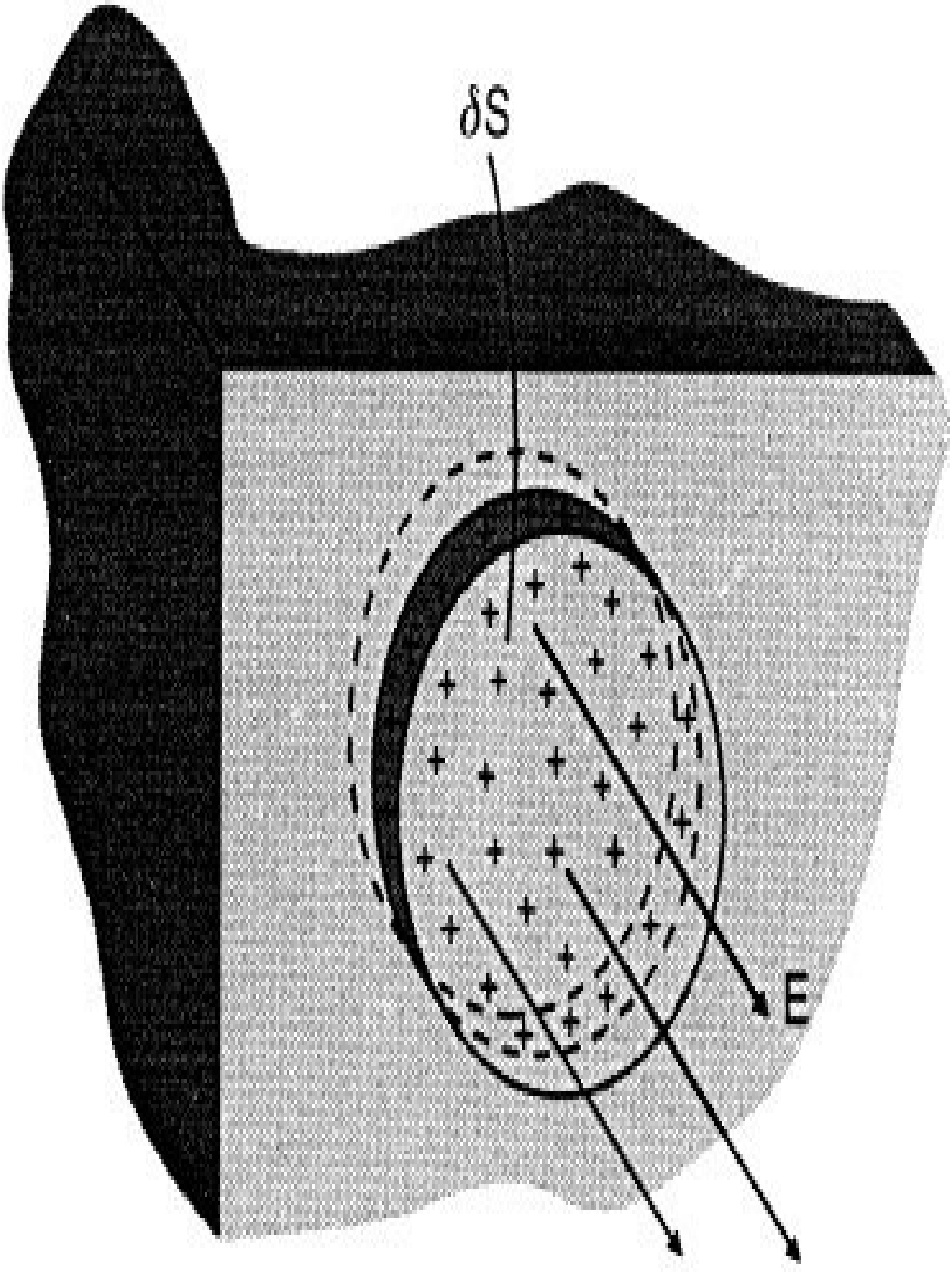
the same field as for a point charge + e located at the origin.

An isolated neutral atom of an inert gas provides another example of spherical symmetry. Since the atom carries no net charge, it follows that the field outside it is exactly zero. Even when they are in the neighbourhood of other atoms, inert gas atoms depart only slightly from spherical symmetry, and the electric fields near them remain small. There are other interactions between atoms besides those due to their electrostatic fields, but it is nevertheless true to say that the feeble chemical activity of inert gases is related to the fact that they are always very nearly spherically symmetrical.

Gauss' law also applies to macroscopic fields and charge densities, since they too are connected by the inverse square law. Figure 1.15 represents a slab of conductor in an external field perpendicular to its surface. Induced surface charge densities $\pm\sigma$ generate an equal and opposite field inside the conductor, where the macroscopic electric field must be zero. Now imagine a disc enclosing an area δS of the surface, as shown in the figure. The flat surfaces of the disc are parallel to the conducting surface while the curved surface is parallel to the external field. The flux $E \cdot \delta S$ out of the flat surface of the disc which is outside the conductor is $E\delta S$ since the vectors E and δS point in the same direction. Inside the conductor the field is zero, and there is no flux out of the other flat surface of the disc; nor is there any flux out of the curved surface, since the field lies along this surface. The total charge enclosed by the disc is $\sigma\delta S$, and application of Gauss' law leads to

(1.16)

Figure 1.15. A disc enclosing some of the induced charge on a conductor.


δS
E

The example of a slab placed perpendicular to an external field is a special case, but it is easy to show that the induced charges are disposed in such a way that the electric field is always perpendicular to the surface. For if there were a component of the field parallel to the surface, conduction electrons would move until they set up a field which exactly cancelled this component. Over a small part of a conducting surface of any shape, the argument given above is valid, and Equation (1.16) holds at each point of the surface, whether or not it is plane.

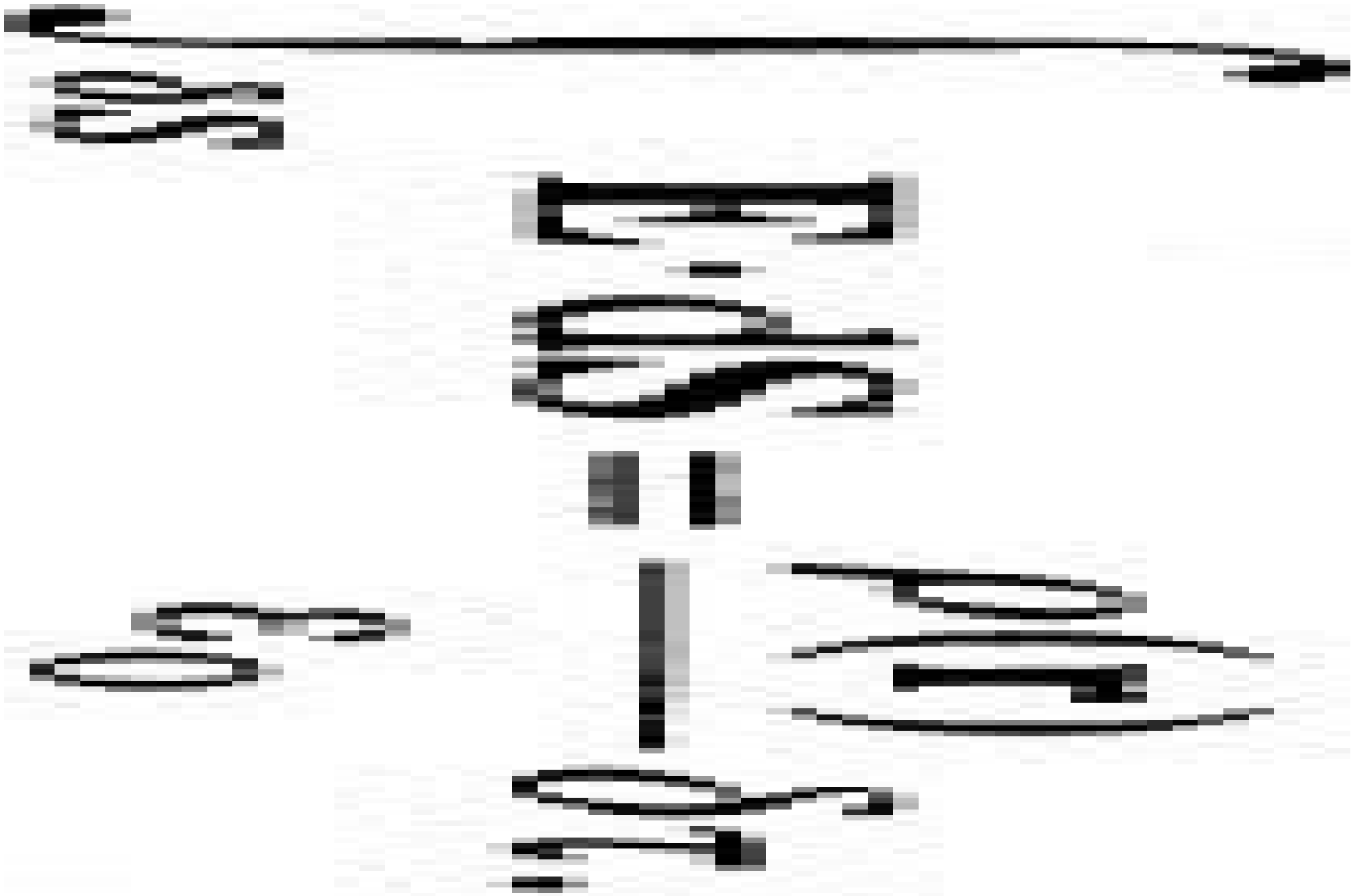
Since macroscopic electric fields are limited to a maximum value of about 10^9 volts m−1, it follows that practically realizable surface charge densities are less than $10^9\varepsilon_0$ 10−2 coulomb m−2. Atomic diameters are about 2 × 10−10m, and the area associated with each atom on the surface of a conductor is about 4 × 10−20 m^2. The maximum surface charge thus corresponds to 4 × 10−22 coulombs, or less than a hundredth of the electronic charge, for each surface atom. The appearance of induced charges represents only a slight perturbation of the outermost layers of atoms in a conductor, and from the macroscopic point of view it is well justified to regard surface charges as infinitesimally thin sheets.

1.4.3 The divergence of a vector field

In the previous section, Gauss' law was given in equation (1.14b) as a relation between two integrals. When the charge density $\rho(r)$ is a smoothly varying function of position, as it is for both macroscopic and atomic charges, Gauss' law can also be expressed as a differential equation. The differential quantity which appears in the equation is called the divergence of the electric field. We shall introduce the divergence in a very general way, using the vector notation, and then find out how to express it in a particular coordinate system.

If we apply Gauss' law to a volume $V = \delta\tau$ small enough for variations in $\rho(r)$ to

be ignored within V, the total charge in V is simply $\rho(r)\delta\tau$, and Equation (1.14b) becomes



Here the integration is over the area δS which encloses $\delta\tau$. In the limit as $\delta\tau$ tends to zero, we have

(1.17)

This equation implies that the quantity on the left-hand side is a scalar function of position, like $\rho(r)$, and that its value does not depend on the shape of the small volume $\delta\tau$. The same limit can be defined for any vector field $F(r)$, and just as for the electric field the result is a scalar field. This limit is the divergence of $F(r)$, and is written div F:

(1.18)

The divergence of a vector field is sometimes written in a different notation as $\nabla\cdot F$. The quantity ∇, called del, is a vector differential operator. Since $\nabla\cdot F$ is the scalar product of the vectors ∇ and F, the notation has the advantage of emphasizing that the divergence of F is a scalar function of position. The use of the operator ∇ is briefly described in Appendix B, but in the main part of the book we shall retain the notation div F.

Although div F has the same value whatever the shape of the small volume $\delta\tau$, its algebraic form does depend on the coordinate system used. Expressions for div F are given in several coordinate systems in Appendix B. It is not difficult to work out the divergence of a vector field in Cartesian coordinates, and we shall do this here for the electric field $E(r)$. In Cartesian coordinates the volume element dr is a rectangular box with sides parallel to the axes and of lengths δx, δy and δz as shown in Figure 1.16. Remembering to take δS always to be the outward normal to the surface, to first order in the small quantities δx, δy and δz

the flux of E out of the sides of the box normal to the x-axis is

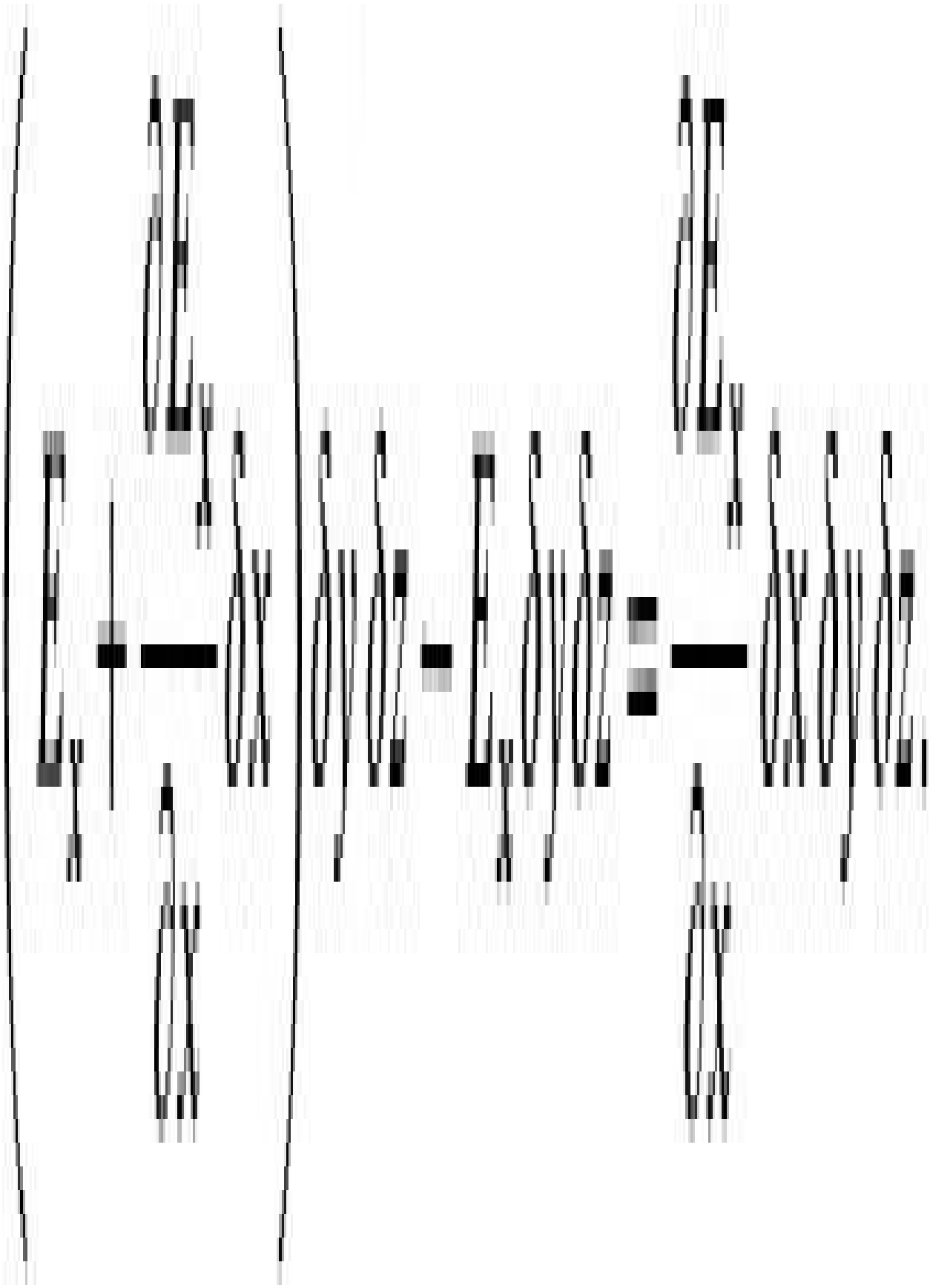
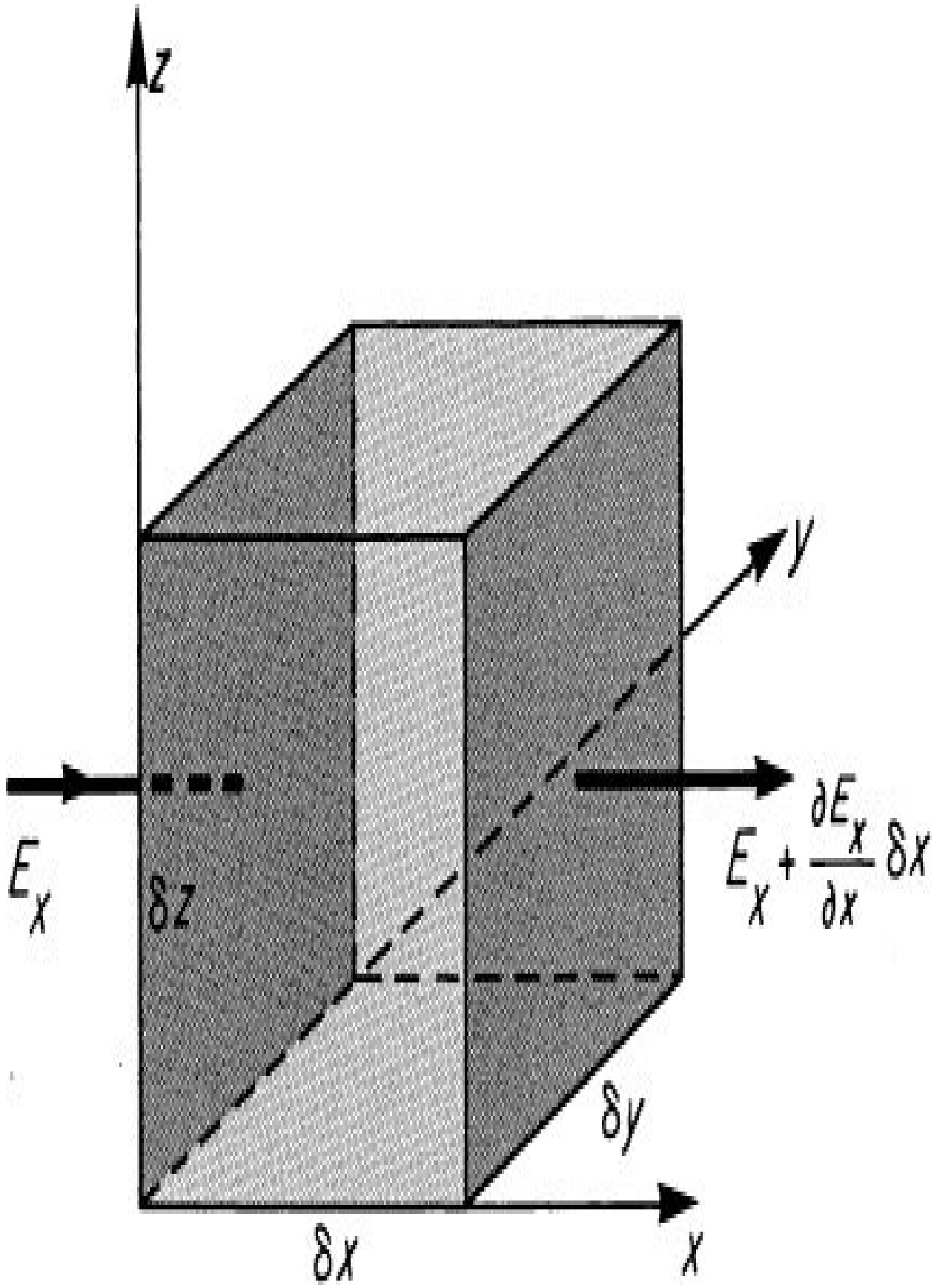
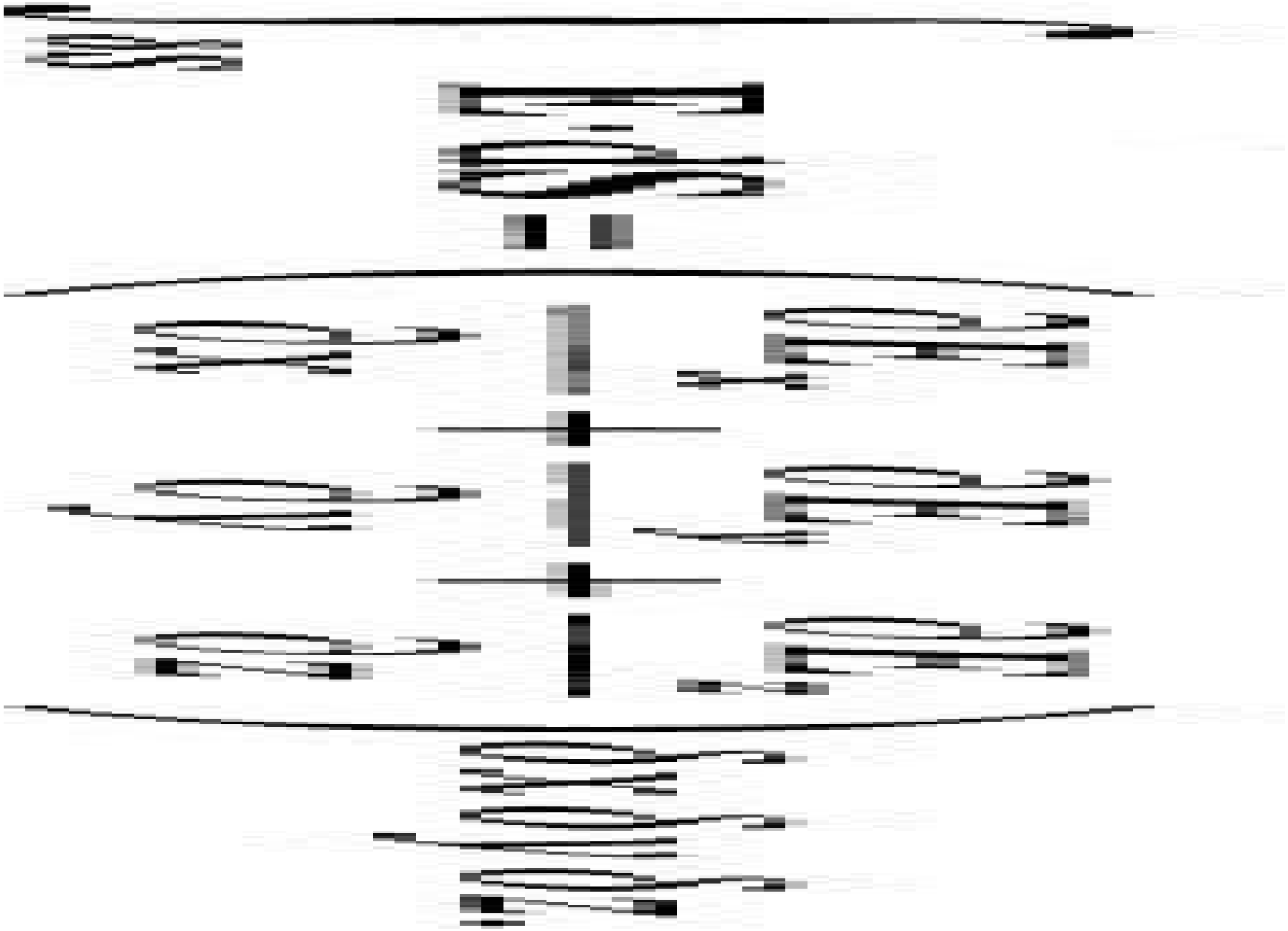


Figure 1.16. The flux out of a rectangular box in a non-uniform field.

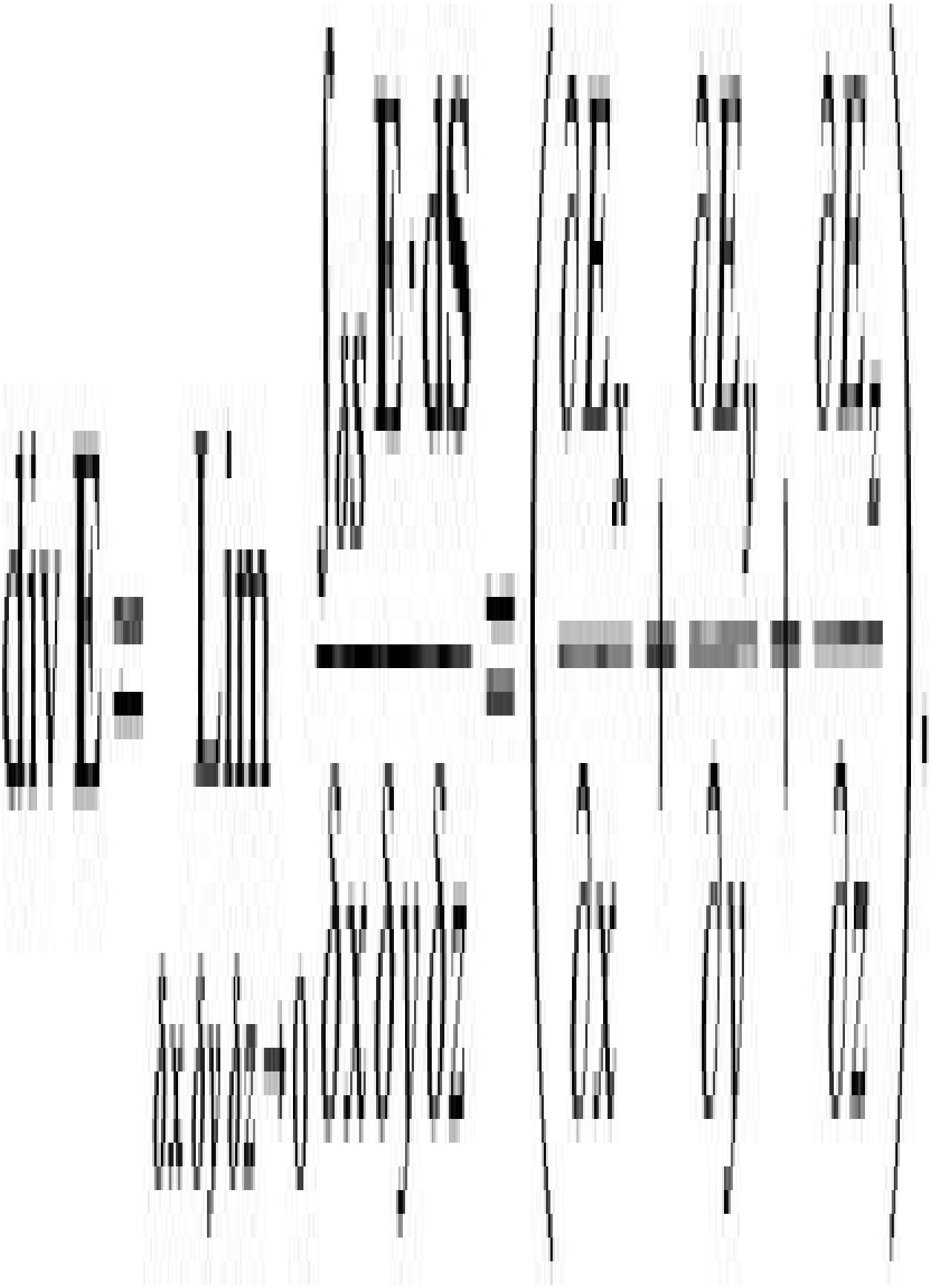

z
y
x
E_x
δz
$E_x + \frac{\partial E_x}{\partial x}\delta x$
δy
δx

There are similar contributions from the sides normal to the y and z axes, and the total flux out of the box is



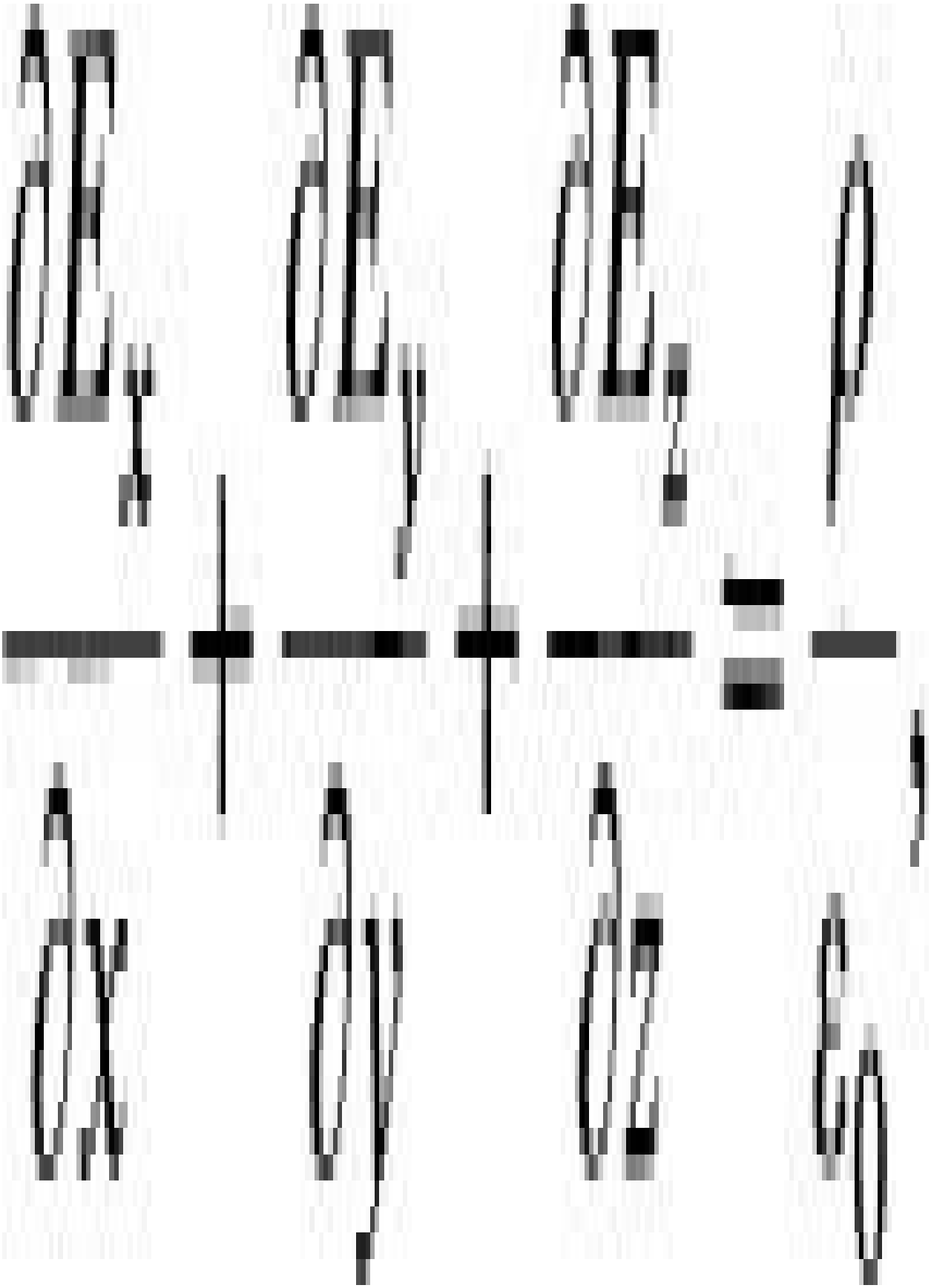
Thus

(1.19)



1.4.4 The differential form of Gauss' law

Comparing Equations (1.17) and (1.19), we find that in Cartesian coordinates



a differential equation relating the electric field to the charge. The equation can be written more generally, without reference to a particular coordinate system, as

(1.20)

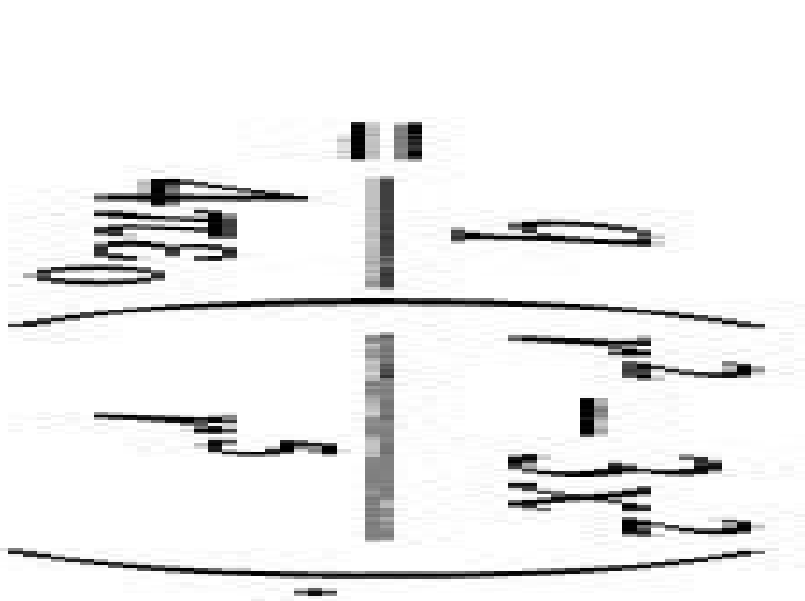
This is the differential form of Gauss' law, which is equivalent to the integral form given in Equation (1.14b). The differential form is the more useful one except when the electric field has a very simple symmetry, and we shall use it to solve some realistic problems in Chapter 3. Equation (1.20) applies to both macroscopic and microscopic electric fields, since in both cases the charge density varies smoothly with position. Although it is rather unrealistic to do so, we often discuss examples in which electric fields are supposed to be generated by fixed point charges. At thepoints where such charges are situated, the charge density is singular, and one must be careful to exclude these points when using the differential form of Gauss' law.

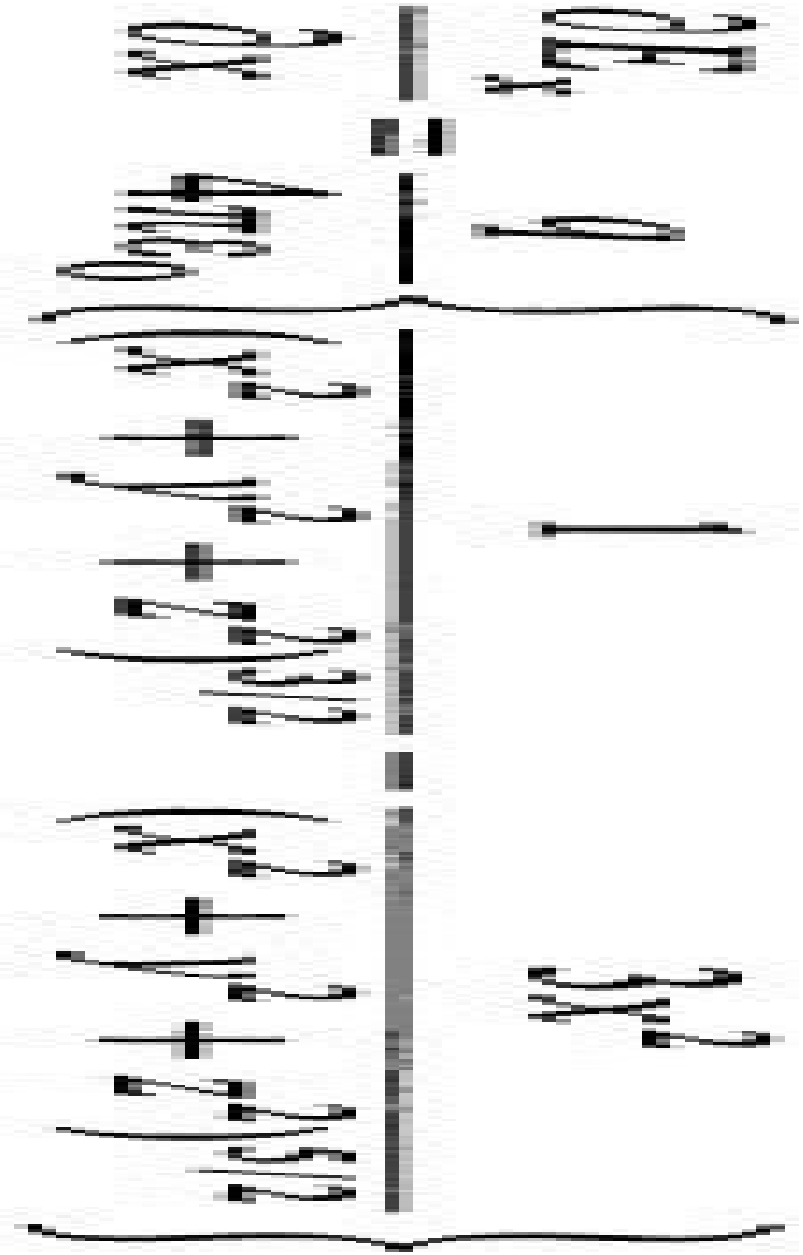
From Equation (1.20) it follows that the divergence of the electric field E is zero in all regions where there is no charge. In the neighbourhood of a point charge, for example, the field is non-uniform, but although the individual terms on the left-hand side of Equation (1.20) may be different from zero, they always sum to zero. To show this, let us place a point charge q at the origin. The position vector r has Cartesian coordinates x, y and z, and the x-component of the electric field at r is

$$E_x = \frac{q}{4\pi\varepsilon_0}\left(\frac{(\mathbf{r})_x}{r^3}\right) = \frac{q}{4\pi\varepsilon_0}\left(\frac{x}{r^3}\right)$$

$$= \frac{qx}{4\pi\varepsilon_0(x^2 + y^2 + z^2)^{3/2}}$$

Differentiating with respect to x (which is possible everywhere except right at the origin, where the field is singular),



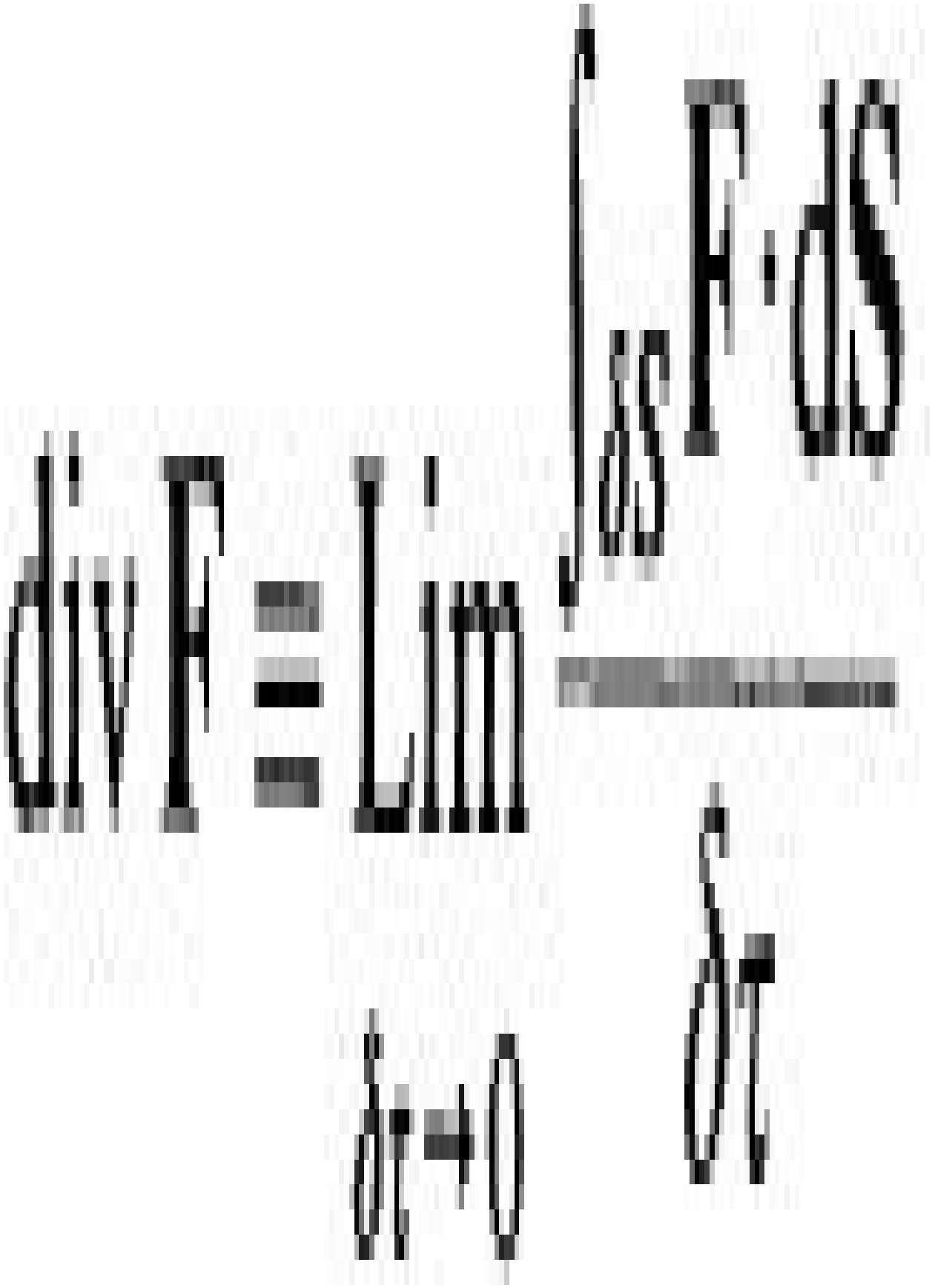


The terms ∂Ey/∂y and ∂Ez/∂z have the same form, with x replaced by y and z in turn. Hence

$$\operatorname{div} \mathbf{E} = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z}$$

$$= \frac{q}{4\pi\varepsilon_0} \left\{ \frac{3r^2 - 3x^2 - 3y^2 - 3z^2}{r^5} \right\} = 0.$$

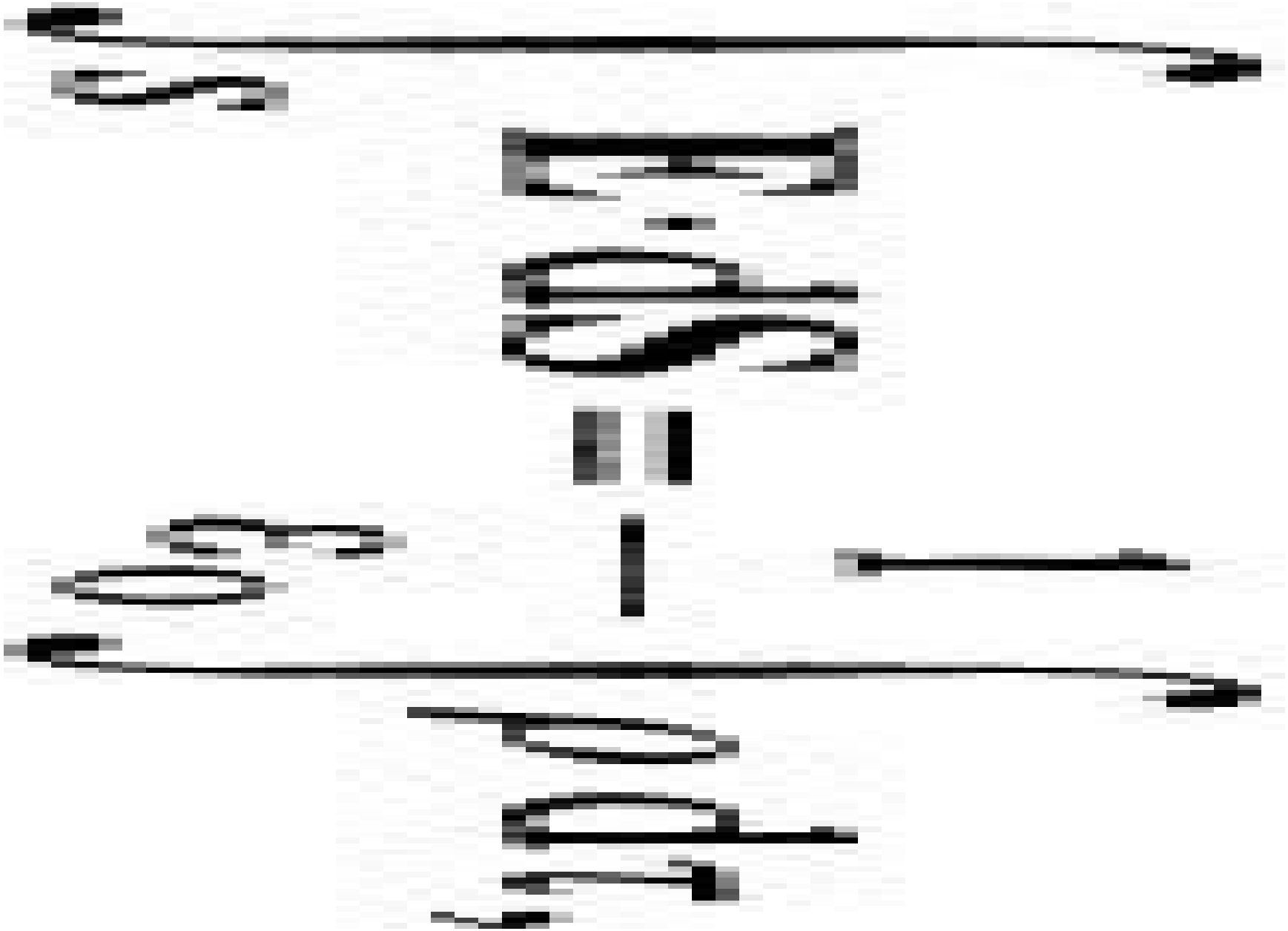
Because Coulomb's law has a straightforward physical meaning, we have chosen to adopt it as the fundamental law of electrostatics. Then we have derived the integral and differential forms of Gauss' law as a consequence of Coulomb's law. But it is also possible to start at the other end of the logical chain, postulating at the very beginning that the electric field is the vector field which is the solution of Equation (1.20). The integral form of Gauss' law can then be deduced very simply. To do this, first return to the definition of divergence given in Equation (1.18):



By integrating this equation over a finite volume, the function div F can be related to the flux of F out of the volume, just as the integral form of Gauss' law relates ρ to the flux of E:

(1.21)

As before, S is the surface enclosing the volume V. The result, an identity which holds whatever the form of F, is often called Gauss' theorem. To avoid confusion with Gauss' law we shall refer to the identity as the divergence theorem. Applying the divergence theorem to the electric field E, and replacing div E by ρ/ε_0, it follows at once that



which is the required result. Now the integral form of Gauss' law implies that Coulomb's law is valid, since if a point charge is placed at the centre of a sphere, only for an inverse square law of force is the flux of the electric field out of the sphere independent of its radius. Coulomb's law, the integral form of Gauss' law and the differential form of Gauss' law are thus equivalent statements, and any one of them may be taken as the fundamental law of electrostatics. Many books on electromagnetism choose Equation (1.20) as the fundamental law; although this equation is more difficult to understand than Coulomb's law, it has the advantage that it is the equation most often used to solve practical problems.

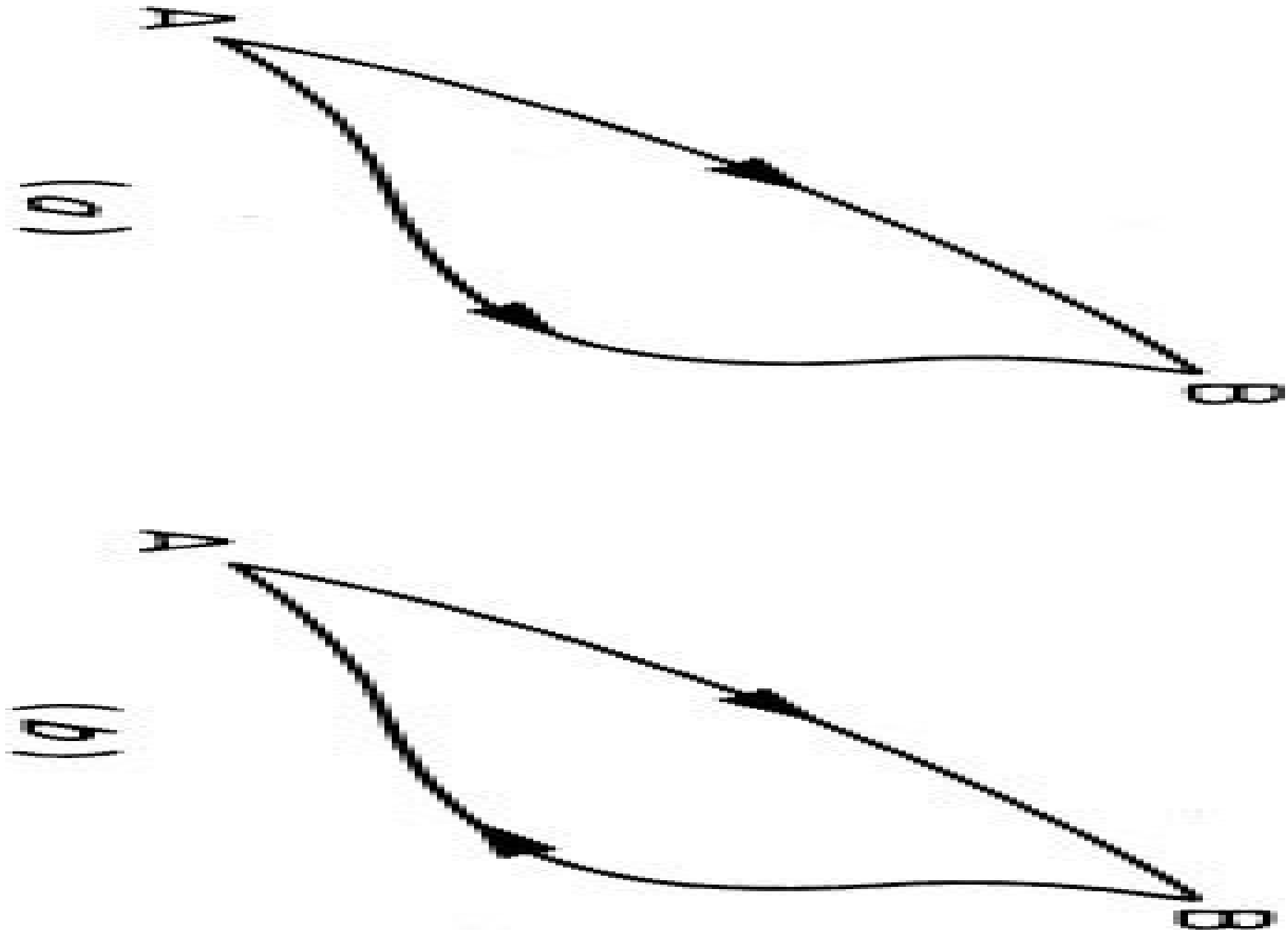
1.5 ELECTROSTATIC ENERGY

Energy changes are associated with alterations in either atomic or macroscopic charge density. The energy liberated in chemical reactions, for example, is almost entirely electrostatic in origin, and is made available by the rearrangement of charges within the atoms undergoing reaction. A flash of lightning, on the other hand, derives its energy from the net charge stored by a thundercloud. Later in section 1.5 we shall meet practical examples of energy changes in both microscopic and macroscopic physics, but we begin by discussing work and energy in terms of idealized point charges, in order to introduce the concept of electrostatic potential.

1.5.1 The electrostatic potential

External work must be done to move a small test charge in opposition to the force it experiences in an electrostatic field. A charge moved in this way acquires potential energy, which may be recovered by allowing the charge to retrace its path under the action of the electrostatic force. We shall show that the net work done when a charge is moved from one place to another does not depend on the path taken by the charge but only on its initial and final positions. For example, if a test charge in an electric field is moved from A to B (Figure 1.17(a)), the work done is the same along both the paths shown. It follows that no work at all is done when a test charge moves along a path which returns it to its original position: in Figure 1.17(b) the work done from A to B is exactly cancelled by the return path from B to A. A force which has this property of doing no work around a closed path (and thus dissipating no energy) is called conservative. In electrostatics the field as well as the force is described as being conservative.

Figure 1.17. Possible paths for a test charge moving between A and B.



To show that the electrostatic field is conservative, it is sufficient to prove that the field around an isolated charge is conservative. The proof can then immediately be extended to all possible fields, since it follows from the principle of superposition that when a test charge is taken round a closed loop in any field whatever, no net amount of work is done on it by any of the other charges in the field. In Figure 1.18 the line joining A and B represents the path along which a test charge qt is moved in the field of a point charge q. If both charges are positive, work is done by the repulsive electrostatic force as the test charge moves from A to B. The work done along the infinitesimal section dl of the path is

qt × (magnitude of the field E) × (component of dl along the direction of E)

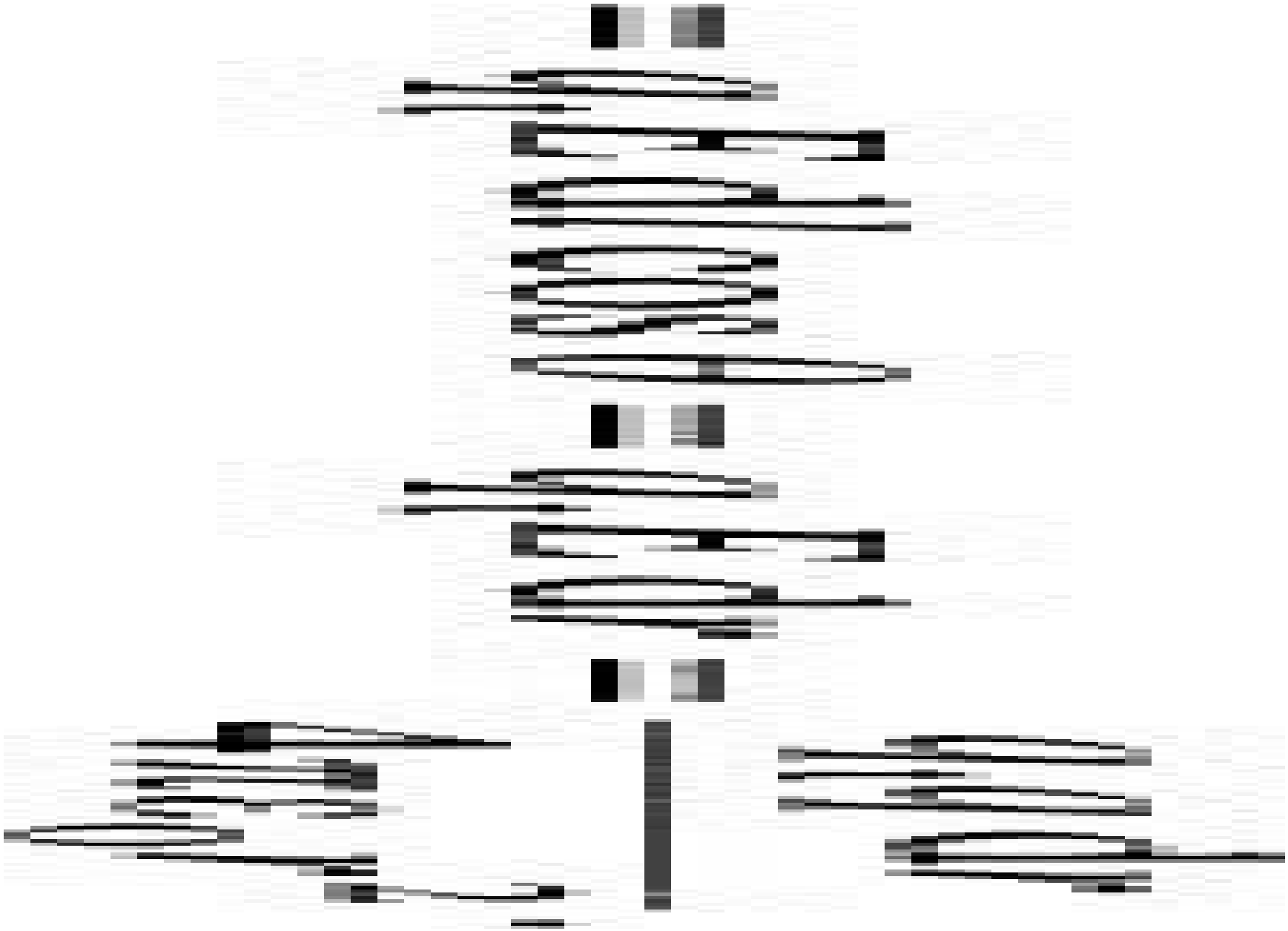
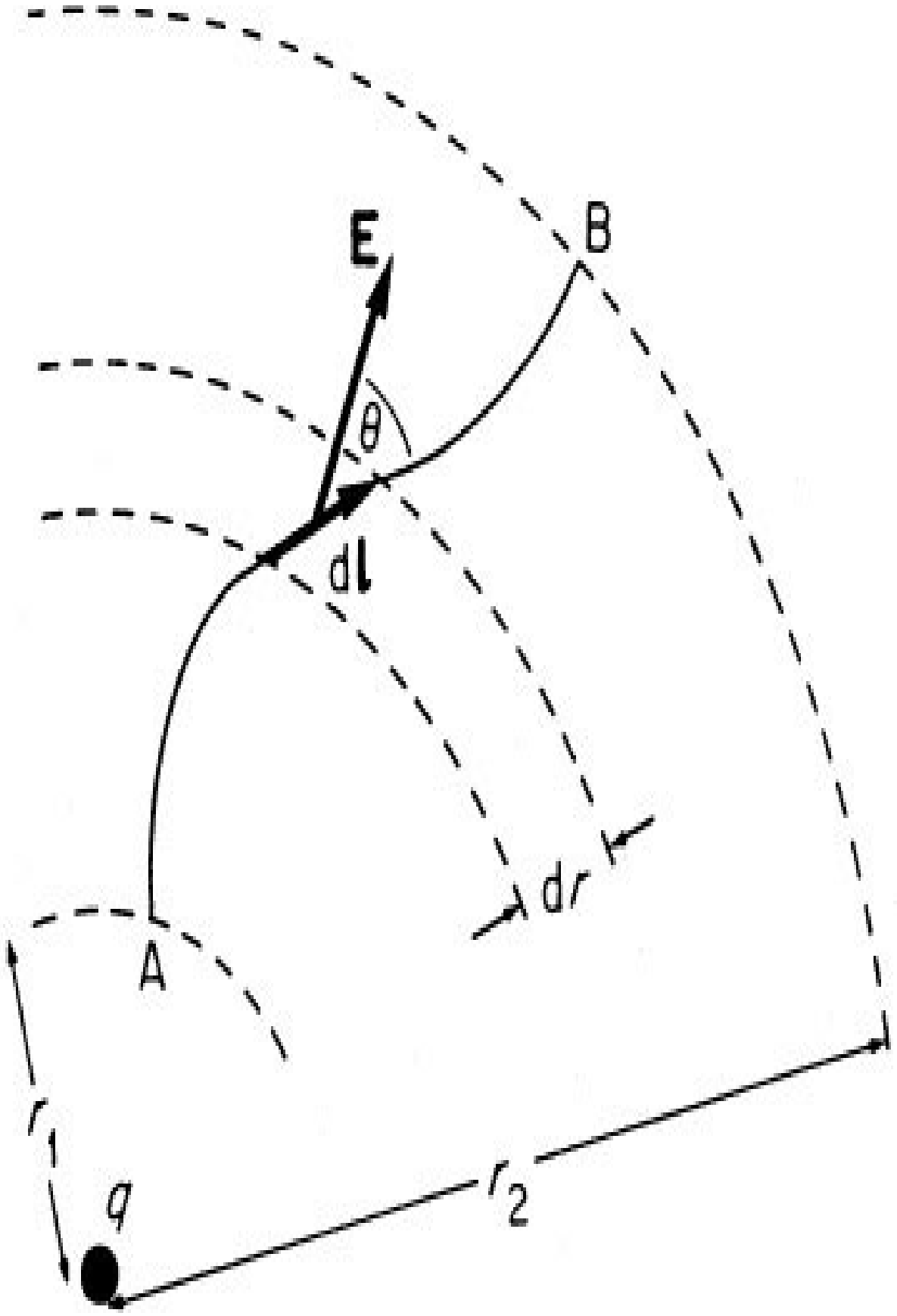


Figure 1.18. Working out the line integral along the path AB.


E
B
θ
dl
dr
A
r1
q
r2

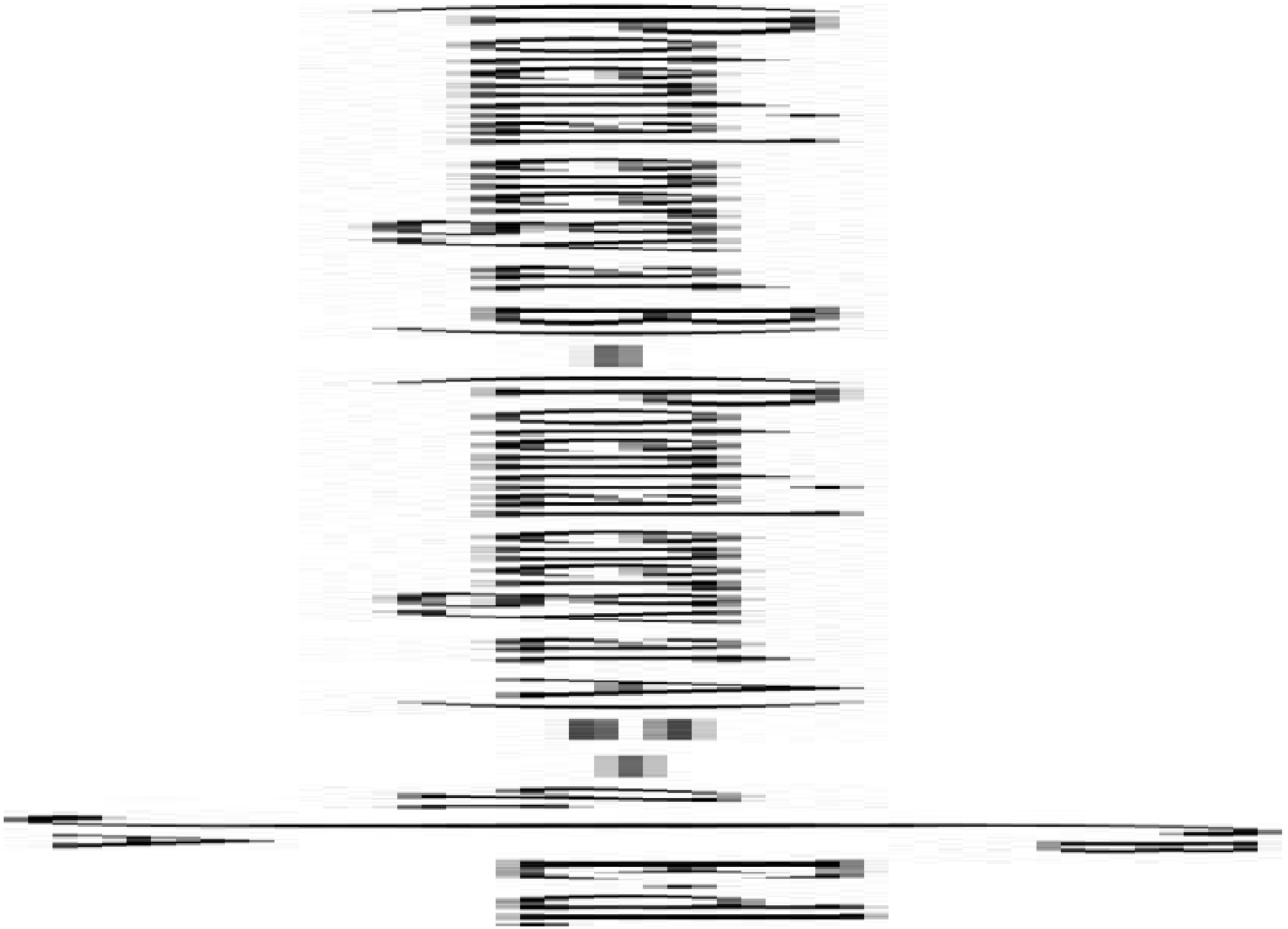
The total work done when the test charge is moved from A to B is thus

(1.22)

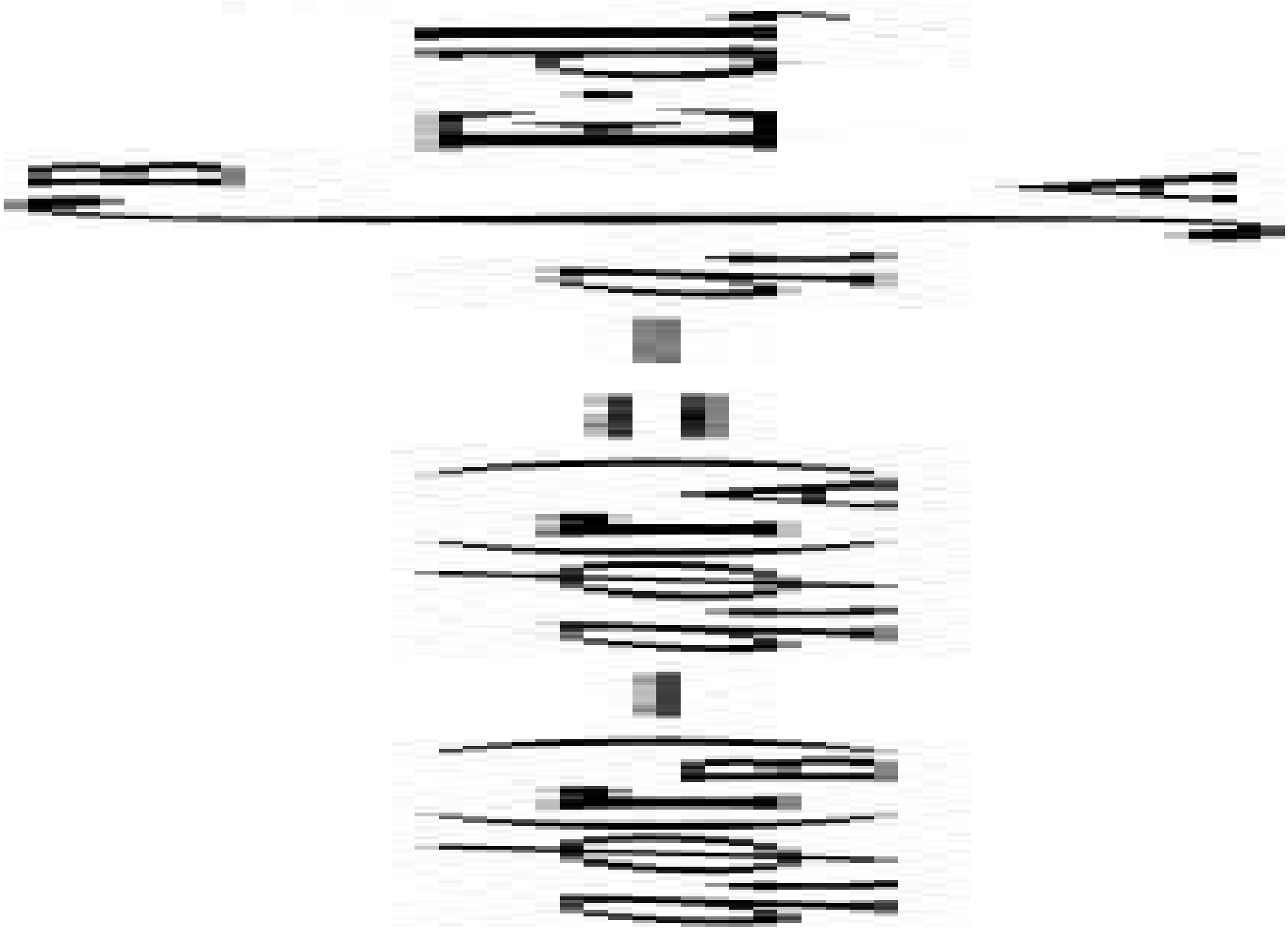
whatever path is used to join A and B. In vector notation the element E dl cos θ occurring in Equation (1.22) is written as E·dl, and the total work done as . The integral is called the line integral of the field E along the path AB. The line integral is independent of the path between A and B for the field in the neighbourhood of a point charge, and we deduce from the principle of superposition that it is independent of the path for all electrostatic fields. Thus, in Figure 1.17(b), the line integral from A to B has the same magnitude but the opposite sign to the line integral along the return path from B to A. The line integral around the closed path from A to B and back again is zero. Since this applies for any closed path in the field, we have proved that the electrostatic field E is conservative. The symbol represents an integral around a closed path, and the mathematical expression of the fact that the electrostatic field is conservative is

(1.23)

Work done by the electrostatic force represents a loss of potential energy of the test charge, for which



Because the electrostatic field is conservative, the potential energy of the test charge depends only on its position, and not on the route taken to get there. In other words, the potential energy is a function of position, and at a point with position vector r it may be written as qtφ(r). The function φ(r) is called the electrostatic potential. Putting rA and rB for the position vectors of the points A and B, the potential energy difference between B and A becomes



or, dividing by qt

(1.24)

Since the electric field E in this equation occurs only in a scalar product, the potential is a scalar function of position, which has a magnitude at any point but no direction. Only differences of potential are defined by Equation (1.24), which is still satisfied if an arbitrary constant is added to $\phi(r)$. The choice of the zero of potential is simply a matter of convenience. For an isolated system of charges, it is usual for the potential at infinity to be chosen as zero. Thus for an isolated point charge situated at the origin, the potential $\phi(r)$ is found from the work done in bringing up a test charge qt from infinity to the point with position vector r. The work done depends only on the magnitude r of the vector r, and is

$$q_1\phi(\mathbf{r}) - q_1\phi(\infty) = -\int_{\infty}^{r} \frac{q_1 q}{4\pi\varepsilon_0 r^2}\,dr$$

$$= \frac{q_1 q}{4\pi\varepsilon_0 r}.$$

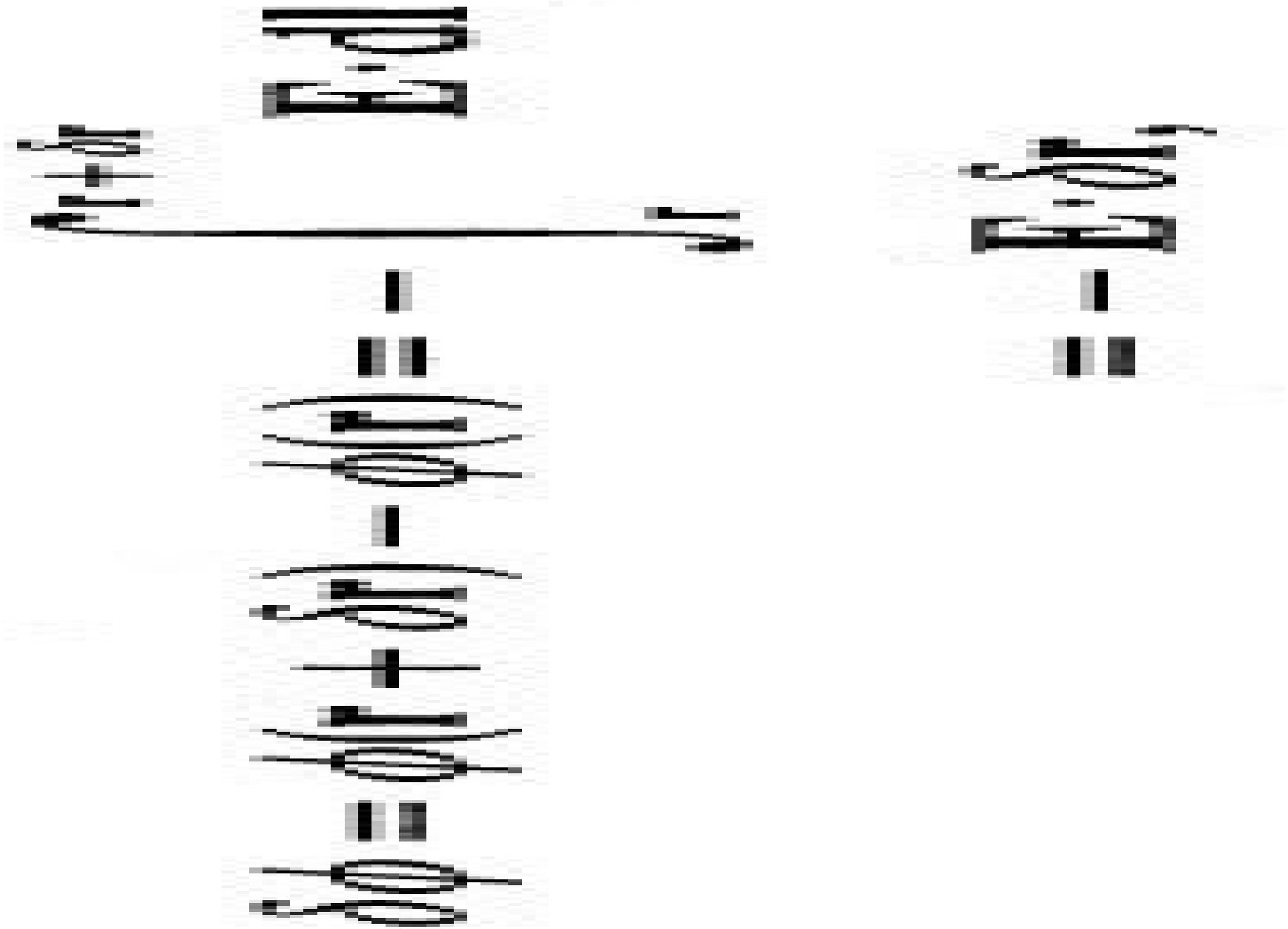
Taking $\phi(\infty)$ to be zero, we have

(1.25)

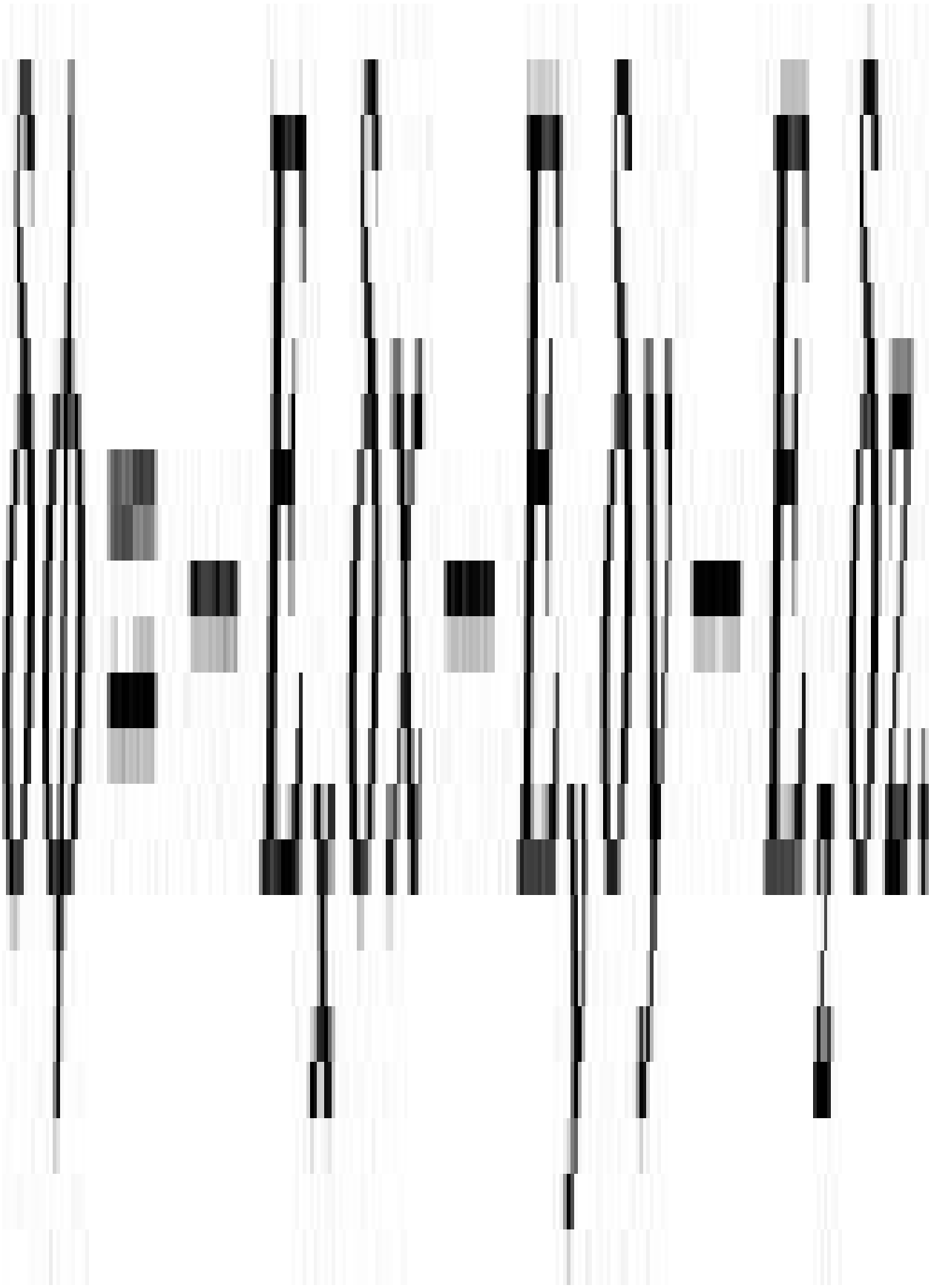
The dimensions of potential are energy per unit charge, and the unit of potential is the volt. One joule of work is done when a charge of one coulomb is moved through a potential difference of one volt.

1.5.2 The electric field as the gradient of the potential

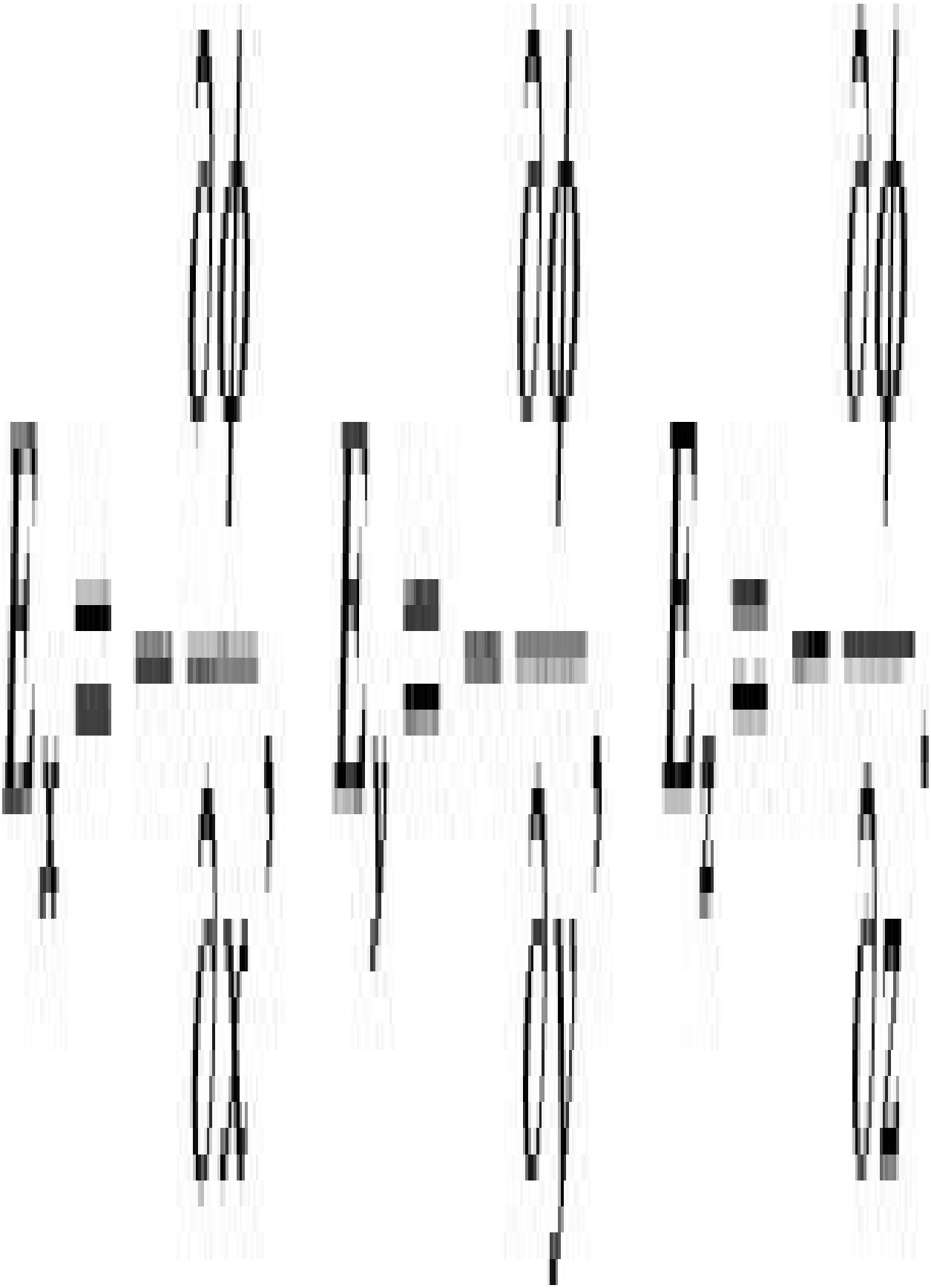
Just as Gauss' theorem could be expressed in integral or differential form, so there is a differential relationship between field and potential which is equivalent to Equation (1.24). The differential relation is found by evaluating the potential difference $\delta\phi$ between the two points a small distance apart, labelled by position vectors r and $(r + \delta r)$:



provided that the interval δr is small enough for E to be considered constant
Working in Cartesian coordinates, the vector δr has components (δx, δy, δz), and $\delta \phi$) becomes



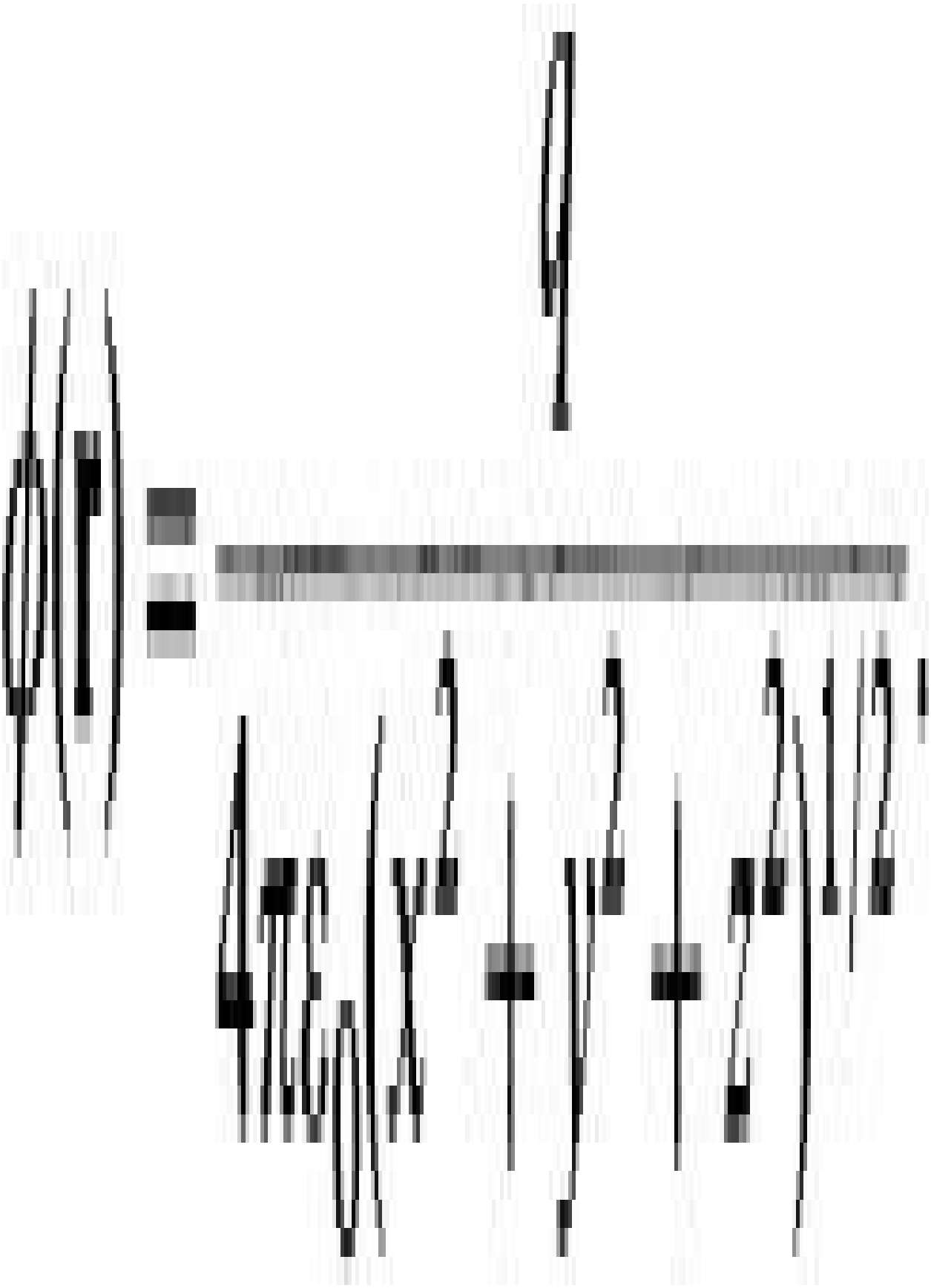
By going to the limit as δx, δy and δz tend to zero we find that the components of the field are partial derivatives of the potentials:



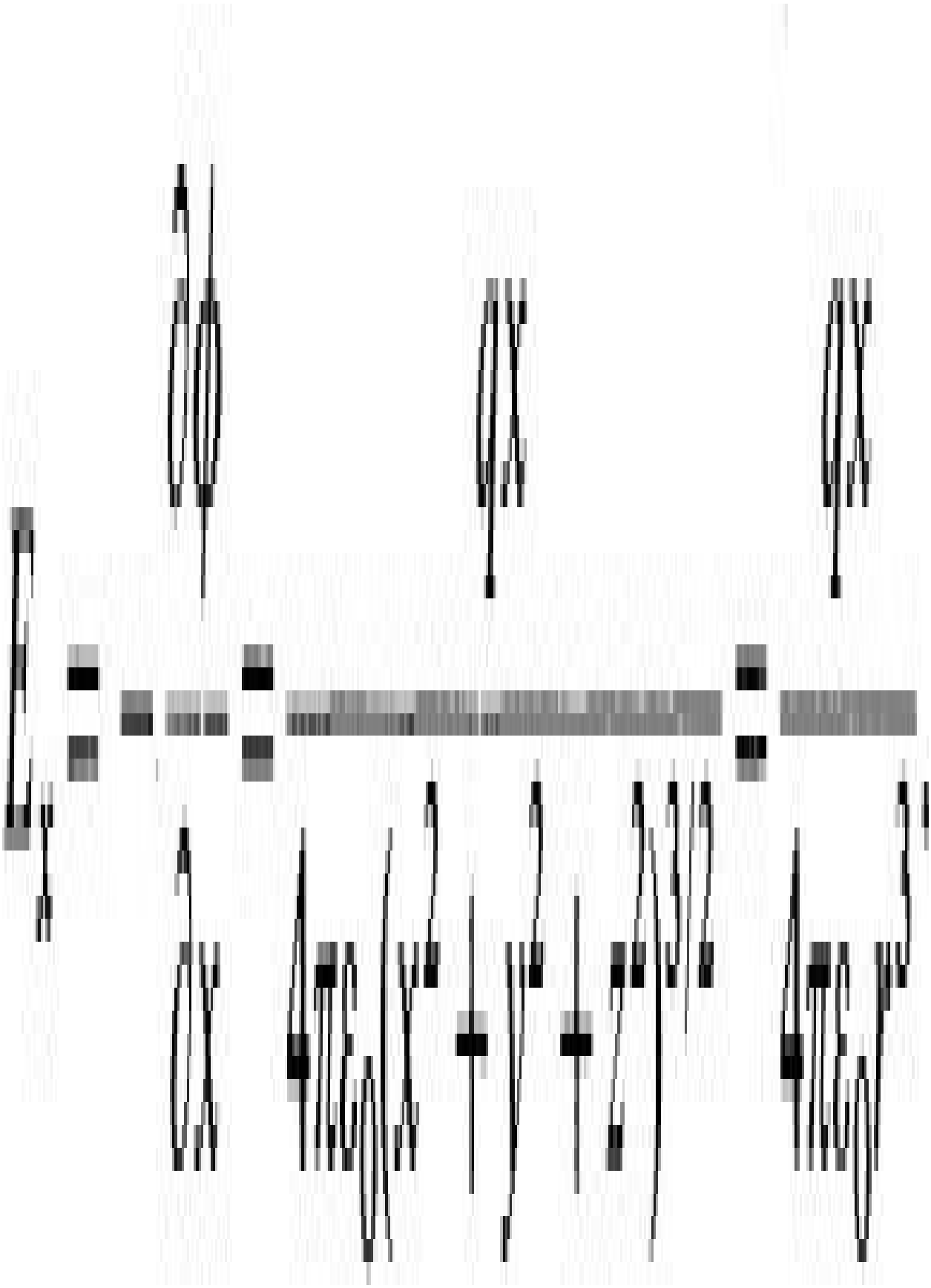
This is a differential relation equivalent to Equation (1.24), written in Cartesian coordinates. The relation may be expressed in a more general form independent of the coordinate system. For any scalar function of position $\psi(r)$, the vector with Cartesian components $\partial\psi/\partial x$, $\partial\psi/\partial y$ and $\partial\psi/\partial z$ is called the gradient of ψ, and is, written as grad ψ. The components of grad ψ for different coordinate systems are given in Appendix B. The gradient of a scalar field is sometimes written as $\nabla\psi$. The vector operator ∇ is the same as the one which occurs in the notation $\nabla.F$ for div F, and the use of this operator applied to scalar fields is also discussed in Appendix B. Using the notation grad ϕ for the gradient of the electrostatic potential, the differential form of Equation (1.24), which is valid in all coordinate systems, is thus

(1.26)

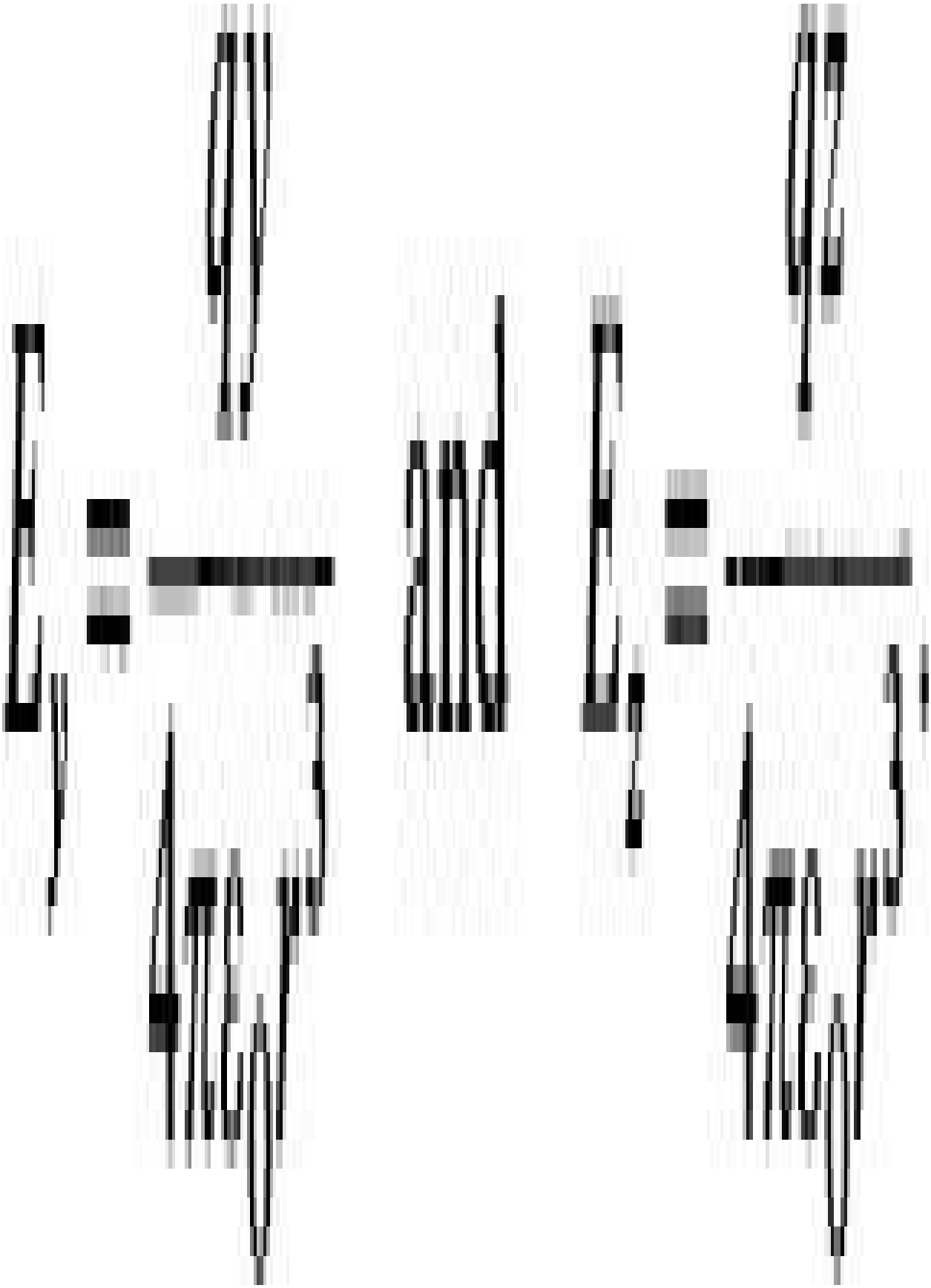
As a check, let us work out the field near a point charge from the potential $\phi(r) = q/4\pi\varepsilon_0 r$ given in Equation (1.25). In Cartesian coordinates, the position vector r has components x, y and z, and



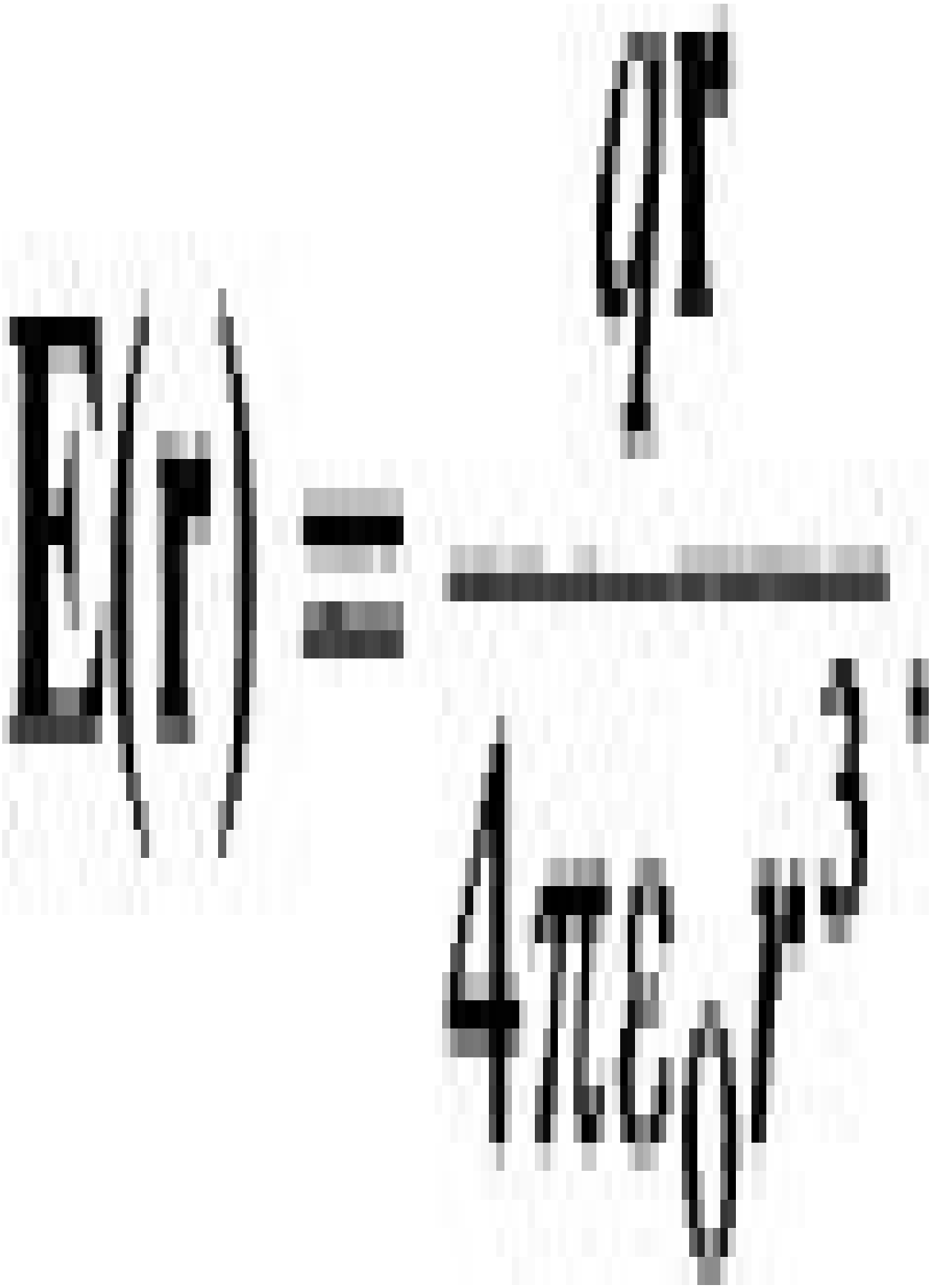
Hence



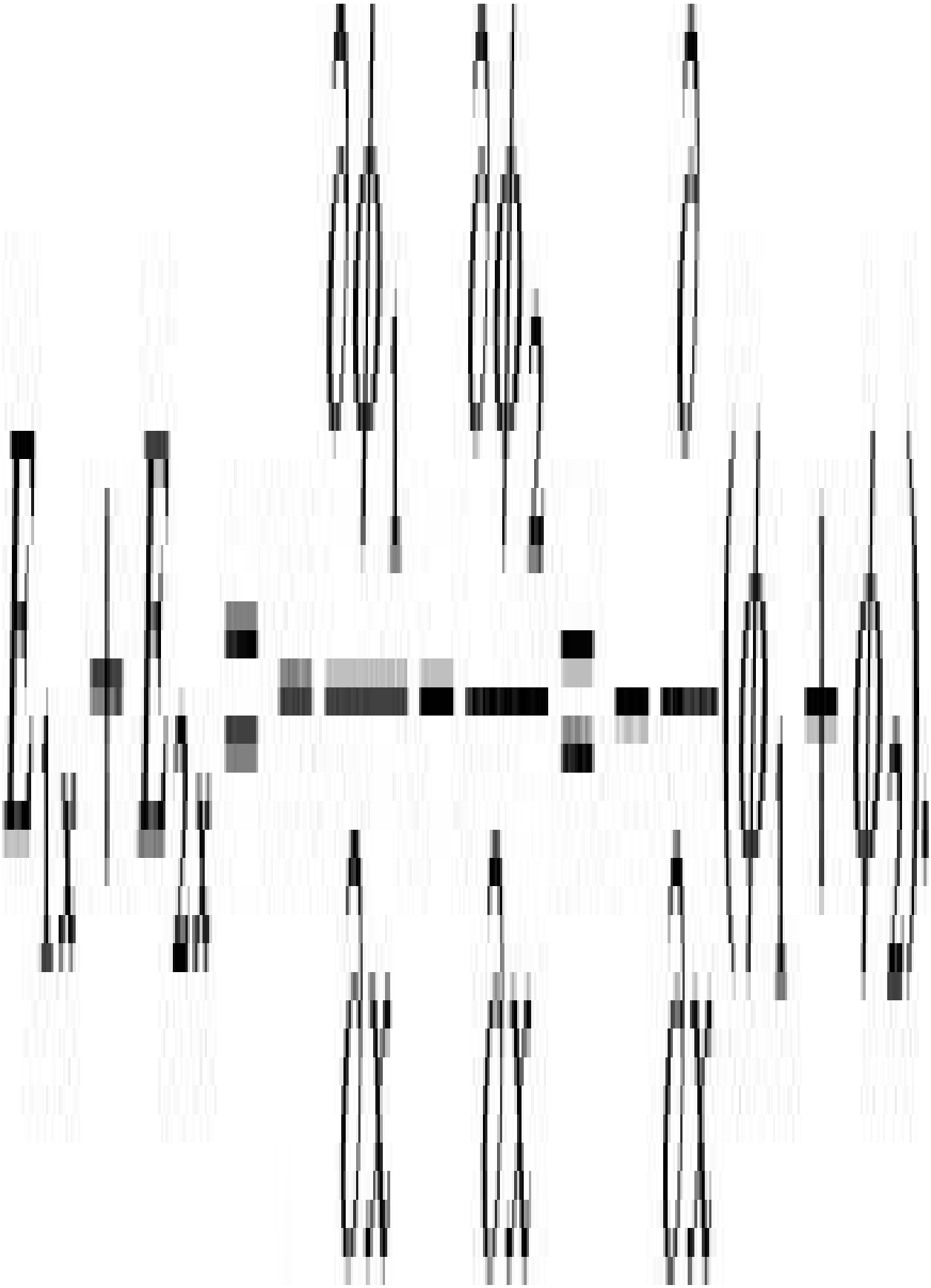
Similarly



These components are proportional to the components of the position vector r, and returning to the vector notation, we recover the correct expression for the field due to a point charge at the origin:



What happens to the potential when two electric fields, generated by different charges, are superimposed? If the fields are E1(r) and E2(r), and the corresponding potentials are $\phi 1(r)$ and $\phi 2(r)$, then the x-component of the field E1 + E2 is

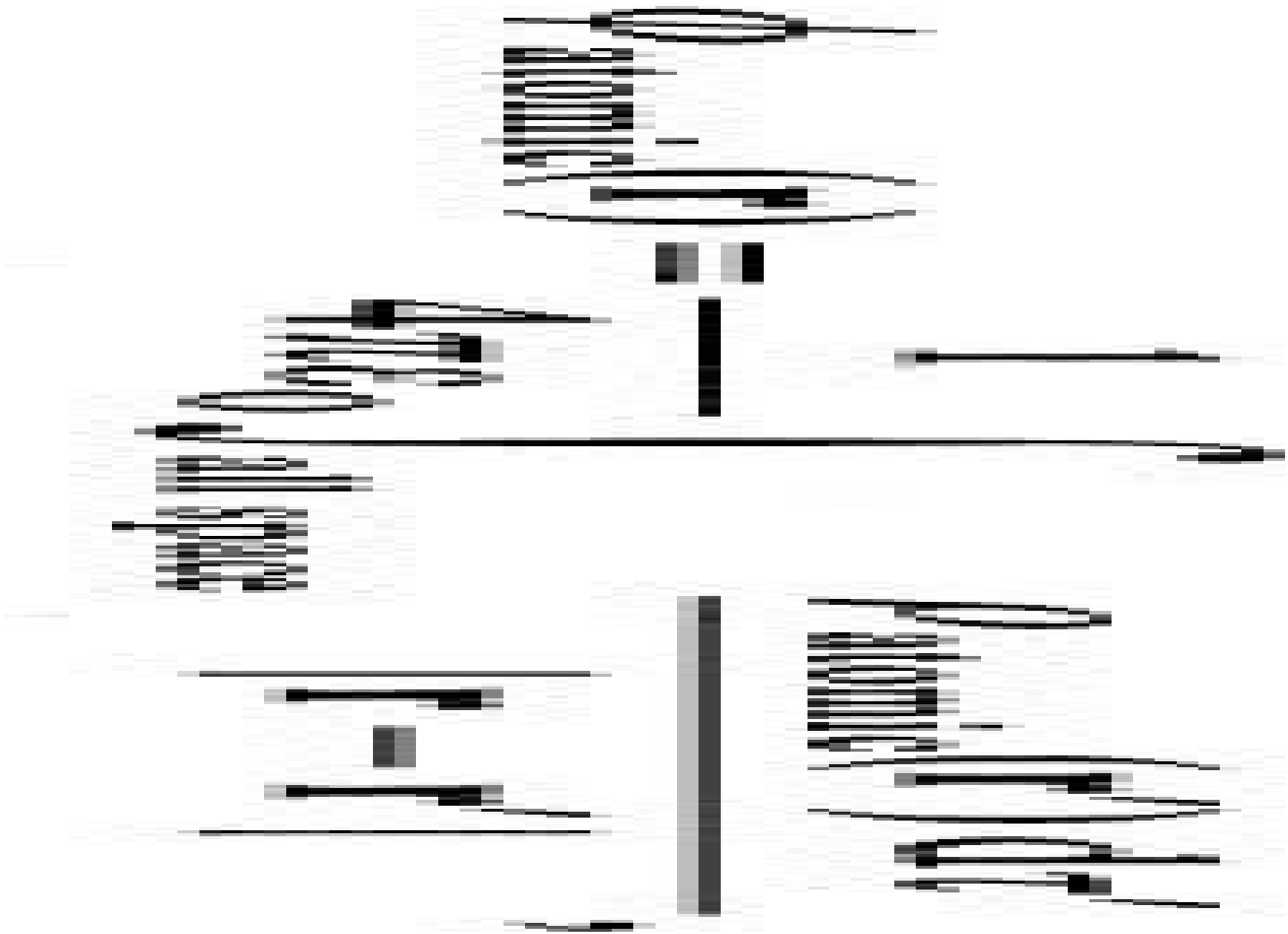


Similar equations hold for the y and z components of the field. The potential for the combined field (E1 + E2) is simply ($\phi 1 + \phi 2$), or in other words, the principle of superposition holds for the potential as well as for the field. The potential at a point P, with position vector r, due to an assembly of charges qi at positions ri, is the sum of the potentials due to each charge separately:

(1.27)

To find the potential due to a continuous charge distribution, the summation in Equation (1.27) is replaced by a volume integral. As with the field, we distinguish between atomic and macroscopic potentials. By comparison with Equations (1.7) and (1.10) for Eatomic and E, the atomic potential is

(1.28)



and the macroscopic potential is

(1.29)

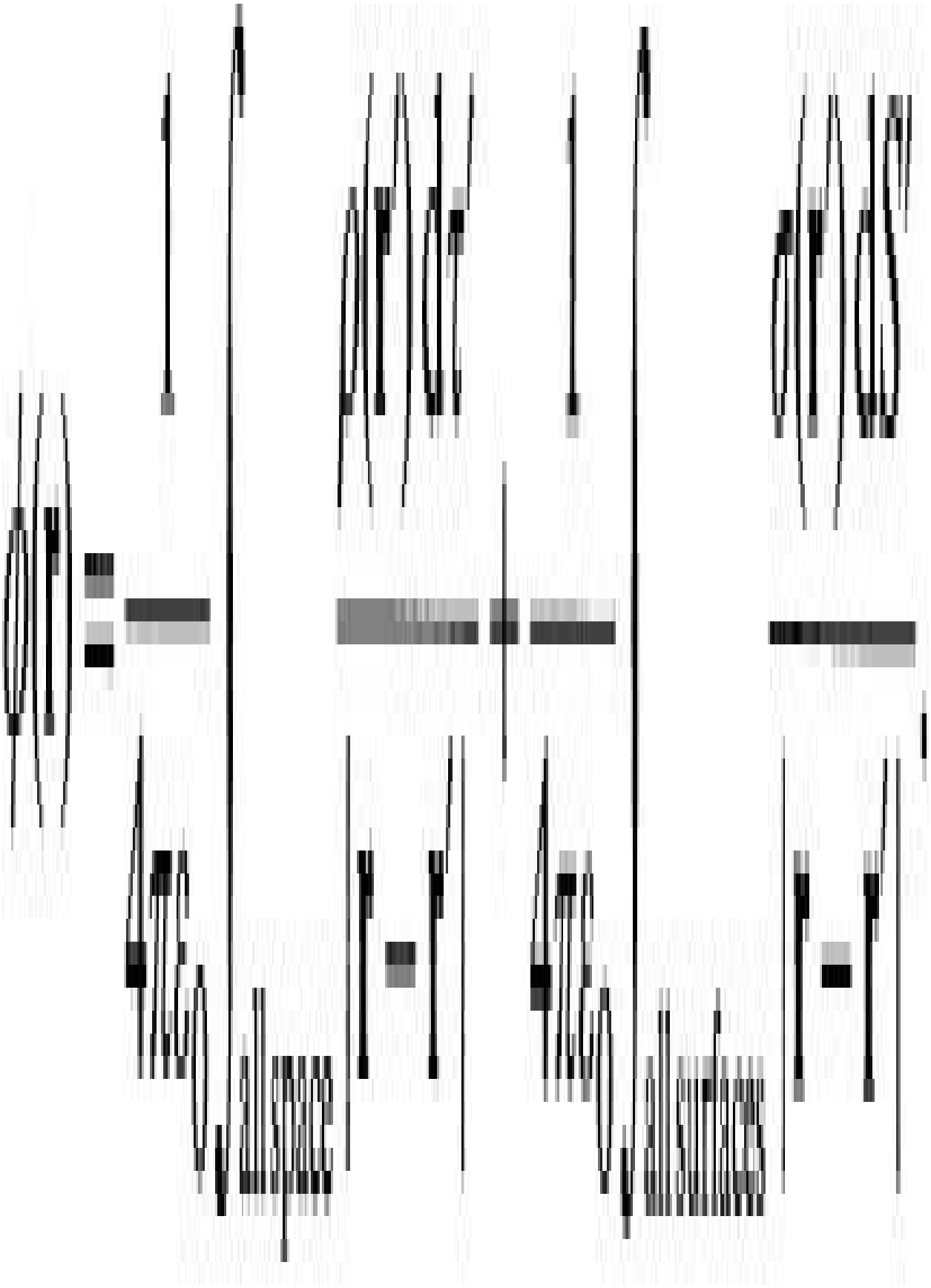
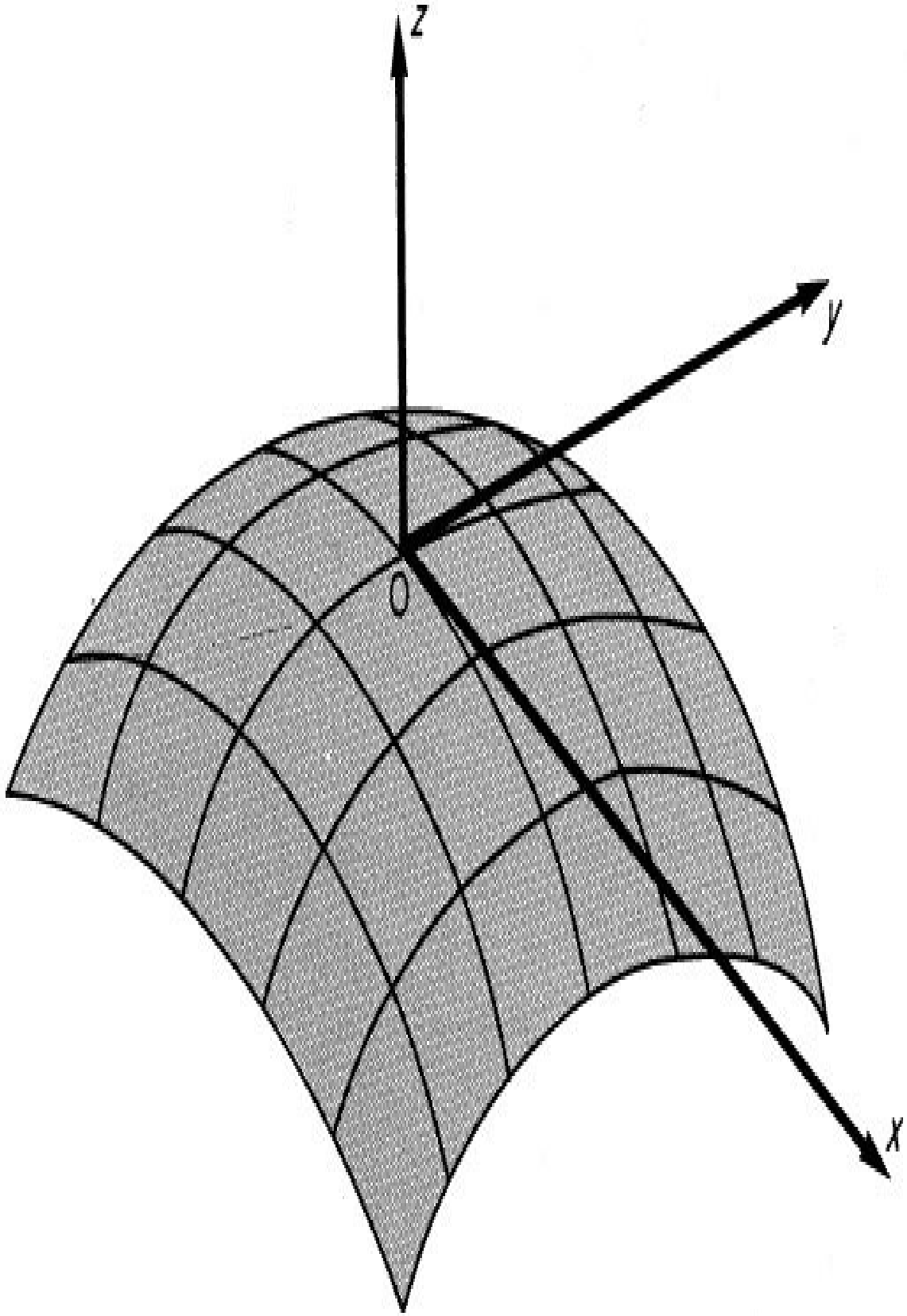


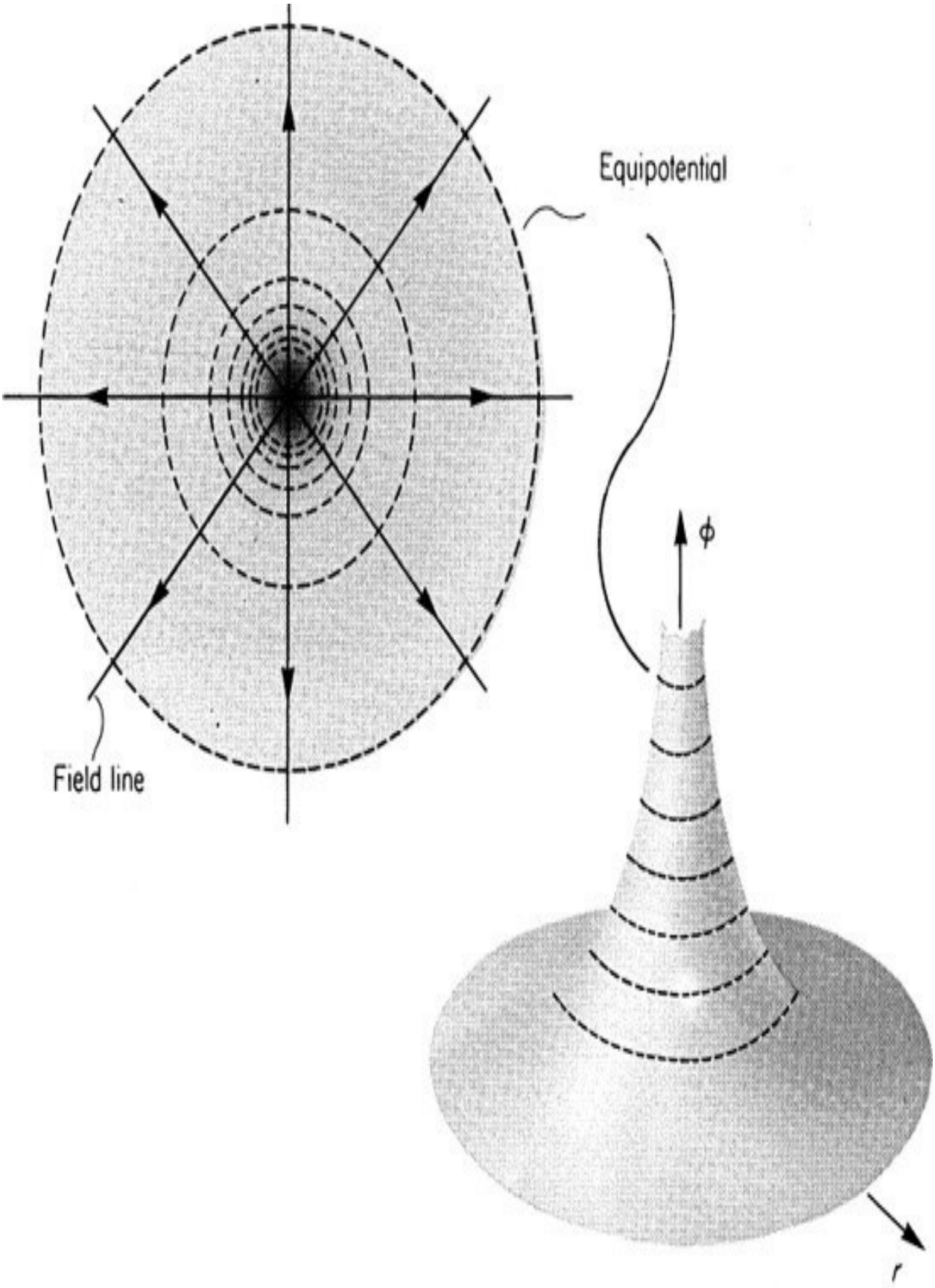
Figure 1.19. The electric field is always normal to equipotential surfaces.


z
y
O
x

Around a point charge + q located at the origin, the potential is $\phi = q/4\pi\varepsilon 0r$. Spheres centred on the origin are surfaces of constant potential, or equipotential surfaces. The lines of force of the electric field are directed outwards from the point charge, and pass perpendicularly through the equipotential surfaces. Lines of force and equipotentials are always perpendicular to one another. To see that this is so, imagine a set of coordinates with origin at the point O on the equipotential surface shown shaded in Figure 1.19. The x and y axes are tangential to the surface at O, and near O, the potential ϕ is then constant in the x and y directions i.e. $\partial\phi/\partial x = \partial\phi/\partial y = 0$. The field is the vector with components $-(\partial\phi/\partial x), -(\partial\phi/\partial y)$ and $-(\partial\phi/\partial z)$ at O the x and y components are zero. The field is therefore in the z-direction normal to the equipotential surface.

Field lines always cross equipotential surfaces perpendicularly, and run in the direction in which the potential decreases most rapidly. Lines of force and equipotentials on a section through an isolated point charge are shown in Figure 1.20 with a sketch illustrating the variation of potential in the plane of the section.

Figure 1.20. The equipotentials and field lines around a point charge.


Equipotential
Field line
ϕ
r

1.5.3 The dipole potential

The spherical equipotential surfaces and radial field lines around a point charge are very simple, but when more than one charge is present the field pattern becomes much more complicated. As an example, let us calculate the potential and electric field of an electric dipole consisting of two equal and opposite point charges. The dipole in Figure 1.21 has two charges ±q displaced from each other in the z-direction, and each at a distance a/2 from the origin. By symmetry the field must have the same pattern in any plane containing the z-axis, but to be definite we shall work in the x−z plane. The potential at the point P lying in this plane with position vector r is the sum of the potentials of each charge separately. Writing r± for the distances from ±q to P, the potential is

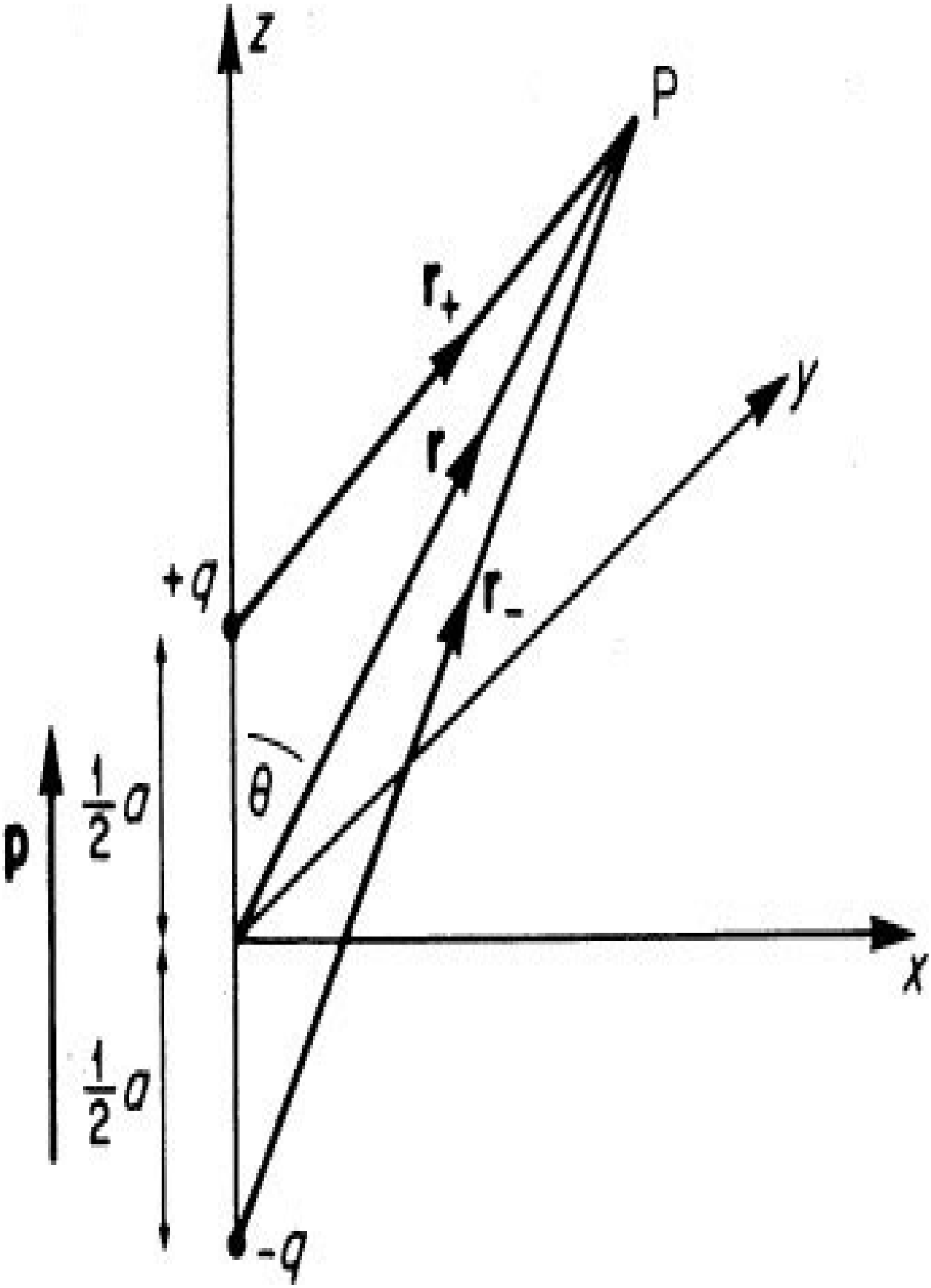
(1.30)

Equipotential surfaces cutting the x−z plane are shown in Figure 1.22, together with the field lines, everywhere perpendicular to the equipotentials. What is the magnitude of the potential in different parts of this diagram? Very close to either of the charges, the potential is almost the same as if the charge were on its own. The equipotential surfaces are nearly spherical and the potential is inversely proportional to the distance from the nearer charge. When r+ becomes comparable to r−, both terms in Equation (1.30) contribute. If θ is the angle between the z-axis and the position vector r, then

$$r_{\pm}^2 = r^2 + \frac{1}{4}a^2 \mp ar\cos\theta$$

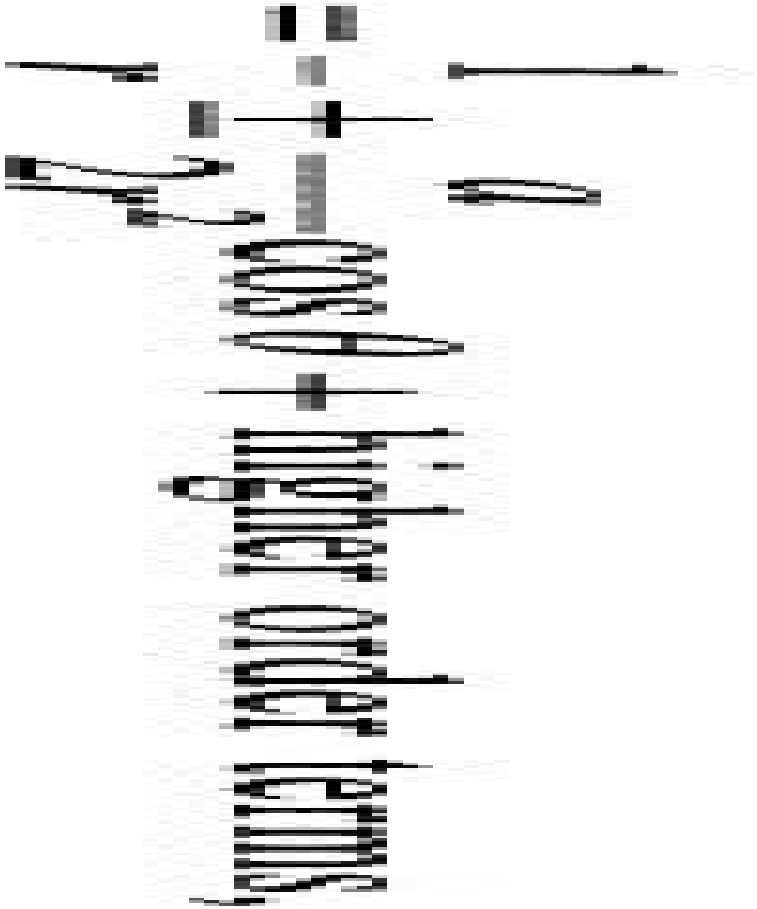
$$= r^2\left(1 + \frac{a^2}{4r^2} \mp \frac{a}{r}\cos\theta\right).$$

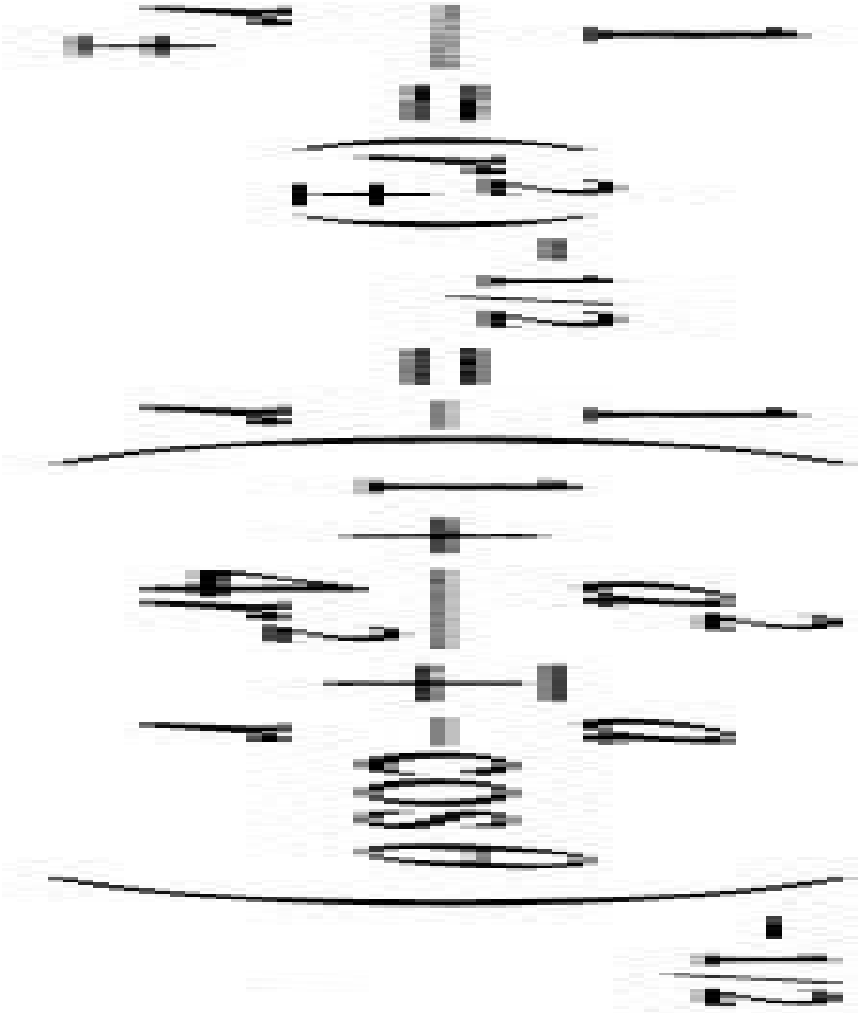
Figure 1.21. The electric dipole.


z
P
r_+
r
y
+q
r_-
θ
p
$\frac{1}{2}a$
$\frac{1}{2}a$
x
-q

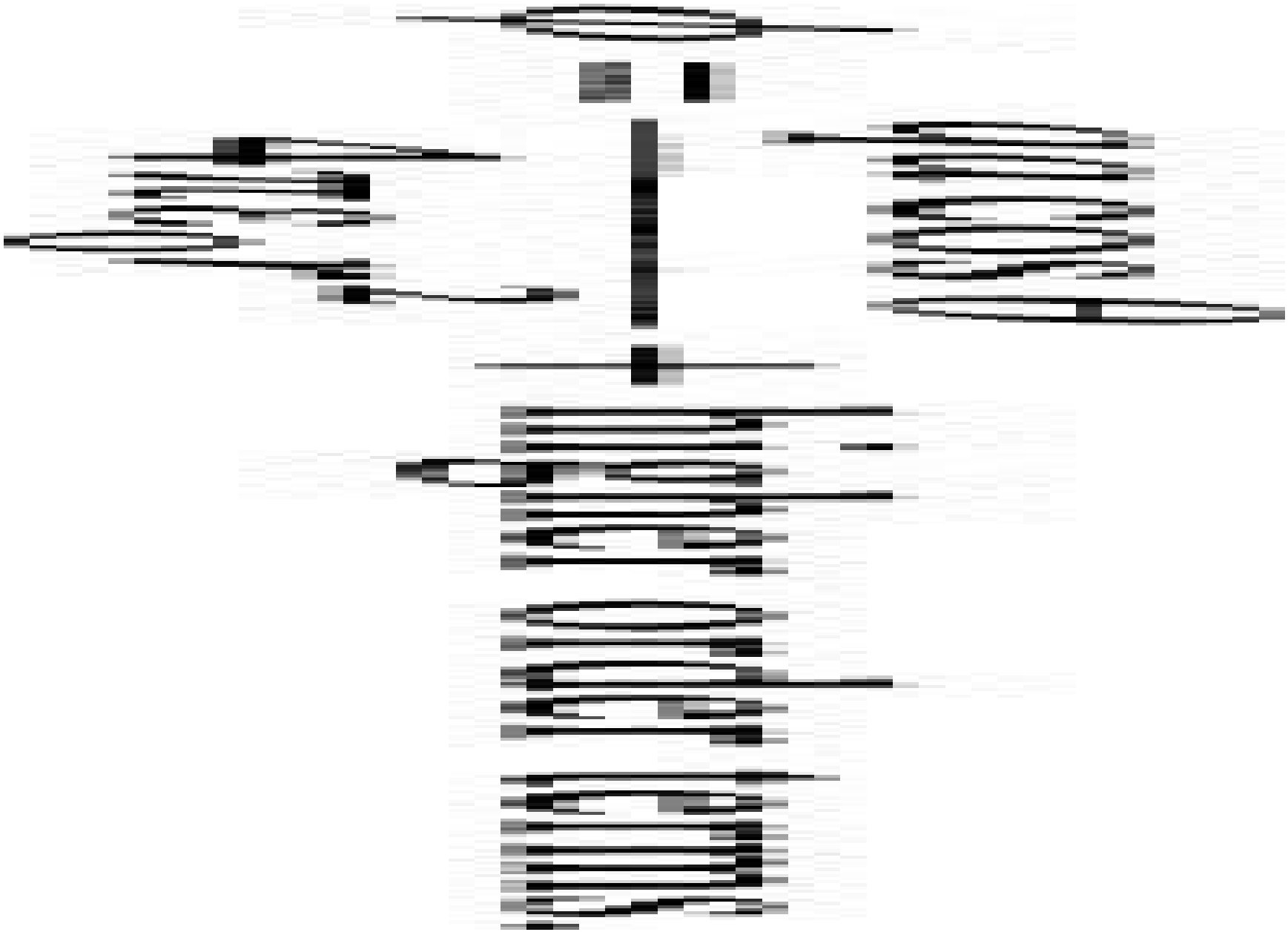
Along the x-axis, where $\cos\theta = 0$, r_+ equals r_-, and the potential is zero (the potential is in fact zero everywhere on the $x-y$ plane at $z = 0$).

At some distance from the dipole, where $r \gg a$, the potential can be expanded in powers of a/r





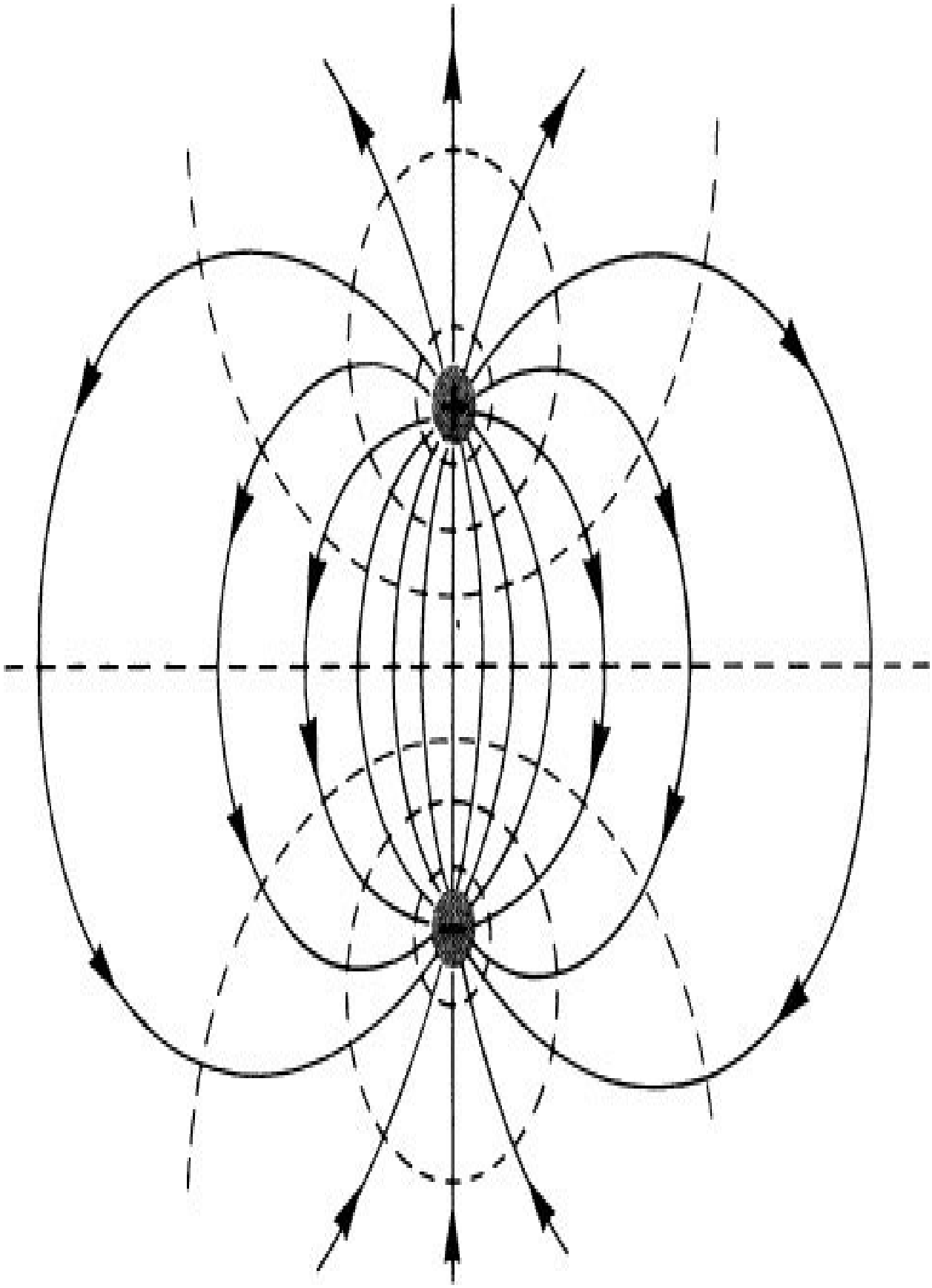
writing out only the term to first order in a/r. Substituting in Equation (1.30) we find



The first term by itself is often referred to as the ‘dipole potential’. It can be regarded as the potential due to a dipole in which the charges ±q have approached within an infinitesimal distance of one another while the product qa of charge and separation remains finite. The ‘dipole potential’ in this sense is exactly proportional to $1/r^2$, but it should be remembered that the potential near a dipole consisting of two separated point charges also includes the higher order terms.

The vector p drawn from the negative to the positive point charge and of magnitude qa is called the dipole moment of the electric dipole. In Figure 1.21 the dipole moment p is directed along the z-axis. The position vector r is at an angle θ to the z-axis, and hence $p.r = qar\cos\theta$. Dropping the higher order terms, the dipole potential can be written as

Figure 1.22. Equipotentials and field lines near an electric dipole.



(1.31)

Notice that this last expression, which is in terms of the vectors p and r, and the magnitude r, makes no reference to a particular choice of coordinate system. The potential is $p.r/4\pi\varepsilon_0 r^3$ even if p is not pointing along the z-axis, and at all points r, whether or not they lie in the x−z plane. In all directions the potential at some distance from the dipole is varying as $1/r^2$, a more rapid fall-off than the 1/r dependence around an isolated point charge. Similarly the electric field around the dipole falls off more rapidly than for a single point charge, and we shall show that the leading term in the expansion in powers of 1/r is proportional to $1/r^3$. We can find the magnitude of the field by calculating the gradient of the potential. In Cartesian coordinates, for a dipole pointing in the z-direction, Equation (1.31) applied to points in the x−z plane has the form

(1.32)

The x- and z-components of the electric field are

(1.33)

and

(1.34)

$$E_z = -\frac{\partial \phi}{\partial z} = -\frac{p}{4\pi\varepsilon_0}\left(\frac{1}{(x^2+z^2)^{3/2}} + \frac{3z^2}{(x^2+z^2)^{5/2}}\right)$$

$$= \frac{p(3\cos^2\theta - 1)}{4\pi\varepsilon_0 r^3}$$

The dipole potential and field are important because they apply to real atoms as well as to idealized point charges. Outside an atom or molecule such as the hydrogen chloride molecule, in which the centres of mass of the positive charge and the negative charge do not coincide, the leading term in the expansion of the potential in powers of 1/r has exactly the same form as Equation (1.31) for the dipole. The higher order terms have different relative magnitudes in the dipole and in the molecule, but at some distance from a molecule only its dipole moment is required in order to specify its contribution to the potential. We shall find in Chapter 2 that the higher order terms can be completely neglected when discussing macroscopic electric fields in matter, and that only the average dipole moment of the constituent atoms and molecules needs to be known.

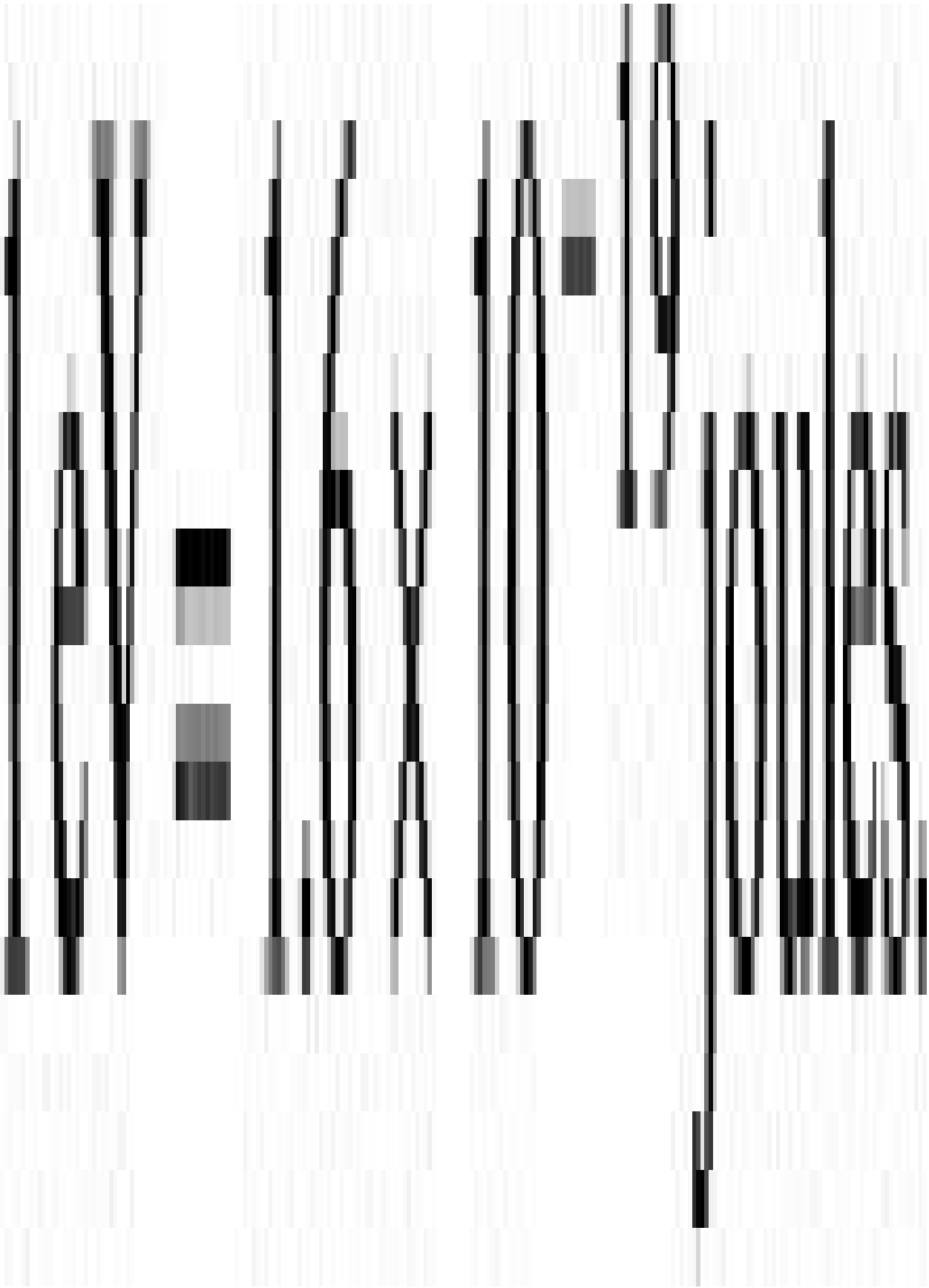
1.5.4 Energy changes associated with the atomic field

The electrostatic potential has very large values near any atomic nucleus. Large amounts of potential energy can therefore be released when the charge carried by a nucleus is redistributed. Such a redistribution occurs in the fission of uranium. The nucleus of the uranium atom splits into two fragments of roughly equal size, which are almost at rest immediately after fission. Both fragments, however, carry a positive charge, and they fly apart because of their electrostatic repulsion. When the fragments are separated by a distance large compared with their radii, they can both be regarded as point charges. If each fragment carries a charge Z times the electronic charge, then the potential energy of the pair of fragments at separation r is $(Ze)^2/4\pi\varepsilon 0r$. It remains quite a good approximation to represent the fragments by point charges even when they are almost touching, although the potential due to one fragment is then by no means constant over the whole extent of the other. If the uranium atom splits into two equal fragments, the charge on each fragment is in fact 46e, and each has a radius of roughly $8 \times 10{-}15$m. When touching, their centres are separated by $1.6 \times 10{-}14$m, and their potential energy is approximately

$$\frac{(Ze)^2}{4\pi\varepsilon_0 r} = \frac{(46 \times 1.6 \times 10^{-19})^2}{4\pi \times 8.85 \times 10^{-12} \times 1.6 \times 10^{-14}}$$

$$= 3.0 \times 10^{-11} \text{ joules.}$$

The joule is an inconveniently large unit of energy when one is discussing single atoms, and in atomic and nuclear physics the usual unit of energy is the electron volt. One electron volt (abbreviated as eV) is the kinetic energy gained by an electron in accelerating through a potential difference of one volt. Since the electronic charge is $1.6 \times 10{-}19$ coulombs,

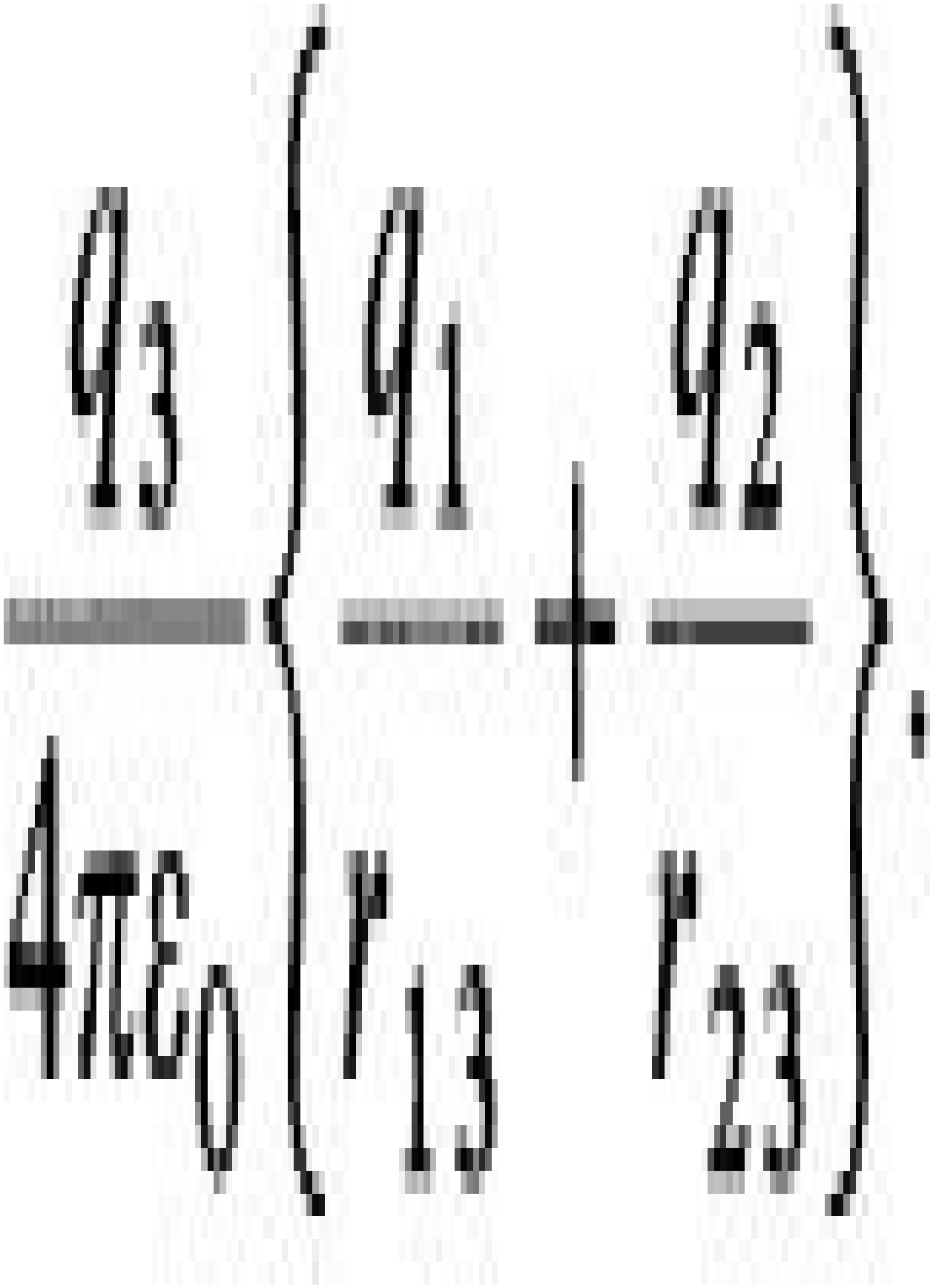


The potential energy of fission fragments in contact, which is equal to their kinetic energy when far apart, is thus about 1.9×10^8 eV. The measured value of kinetic energy for the fragments formed by fission of uranium is 1.7×10^8eV: the difference from our estimate occurs mainly because the fragments are not in fact spherical. This is a very large amount of energy, as is to be expected since electrostatic forces are very strong near the atomic nucleus. When one gram atom of uranium (about half a pound) undergoes fission in a reactor, the energy released as fragment kinetic energy is $6 \times 10^{23} \times 3 \times 10{-}11\ 2 \times 10^{13}$ joules, enough to operate a 1000 megawatt power station for an hour or so.

Another example of energy changes associated with the atomic field is the energy of chemical reactions, which is derived from the rearrangement of electrons in atoms and molecules. Atomic radii are about ten thousand times as great as nuclear radii, and because electrons are on the average further apart than particles in the nucleus, their electrostatic interactions are much weaker. Chemical energy changes are millions of times smaller than the energy released in fission. An example of a chemical energy change which can be estimated easily is the heat of formation of an ionic crystal such as sodium chloride. The crystal consists of a regular lattice of positively charged sodium ions and negatively charged chlorine ions. The ions are formed by prising away one electron from each sodium atom and transferring it to a chlorine atom. Energy is required to effect the transfer, amounting to 1.5eV per ion pair. However, more than 1.5eV is available because of the mutual attraction of the oppositely charged sodium and chlorine ions. Energy is therefore released in formation of the crystal from its constituent atoms; this energy is called the heat of formation. To calculate the heat of formation of the ionic crystal we use the same approximation as for the fission fragments, and represent each ion by a point charge at its centre. All that remains to be done is to apply the principle of superposition to find the potential energy of an assembly of many ions.

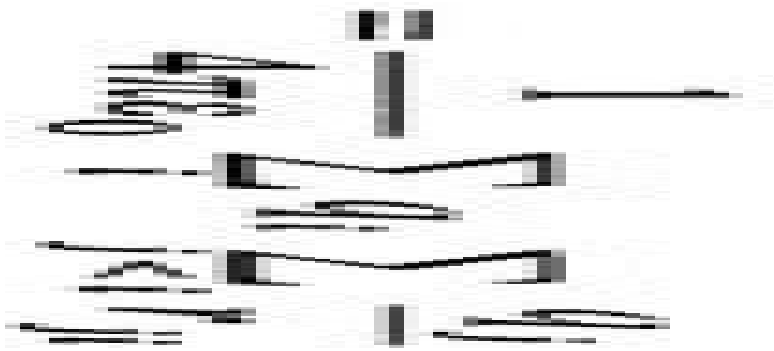
Imagine an isolated charge q1 When a second charge q2 is brought up to a distance r12 from the first, the potential energy of the system is q1q2/4πε0r12. To bring up a third charge q3 requires work to be done against the fields of both

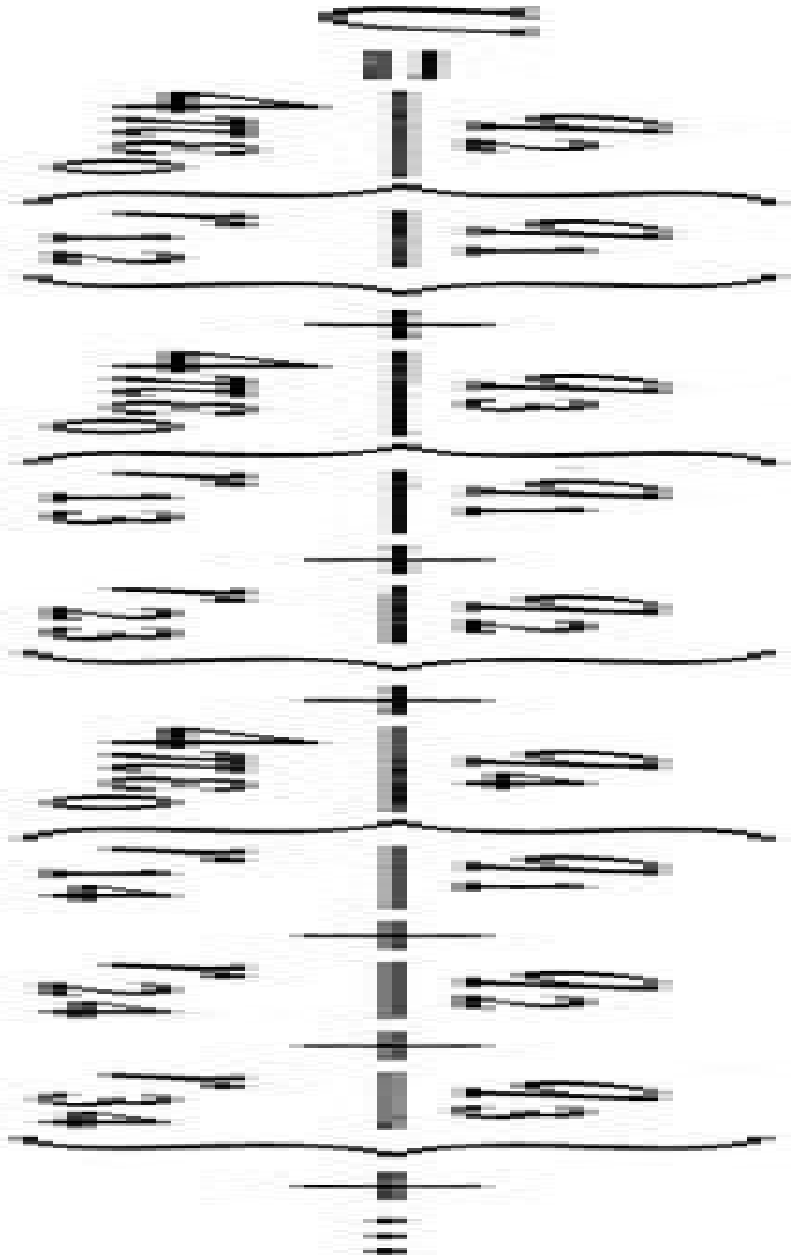
q1 and q2. If the final position of q3 is r13 from q1 and r23 from q2, the additional potential energy is



By building up the whole assembly in this way, a single charge at a time, we see that the total potential energy of the assembly is

(1.35)



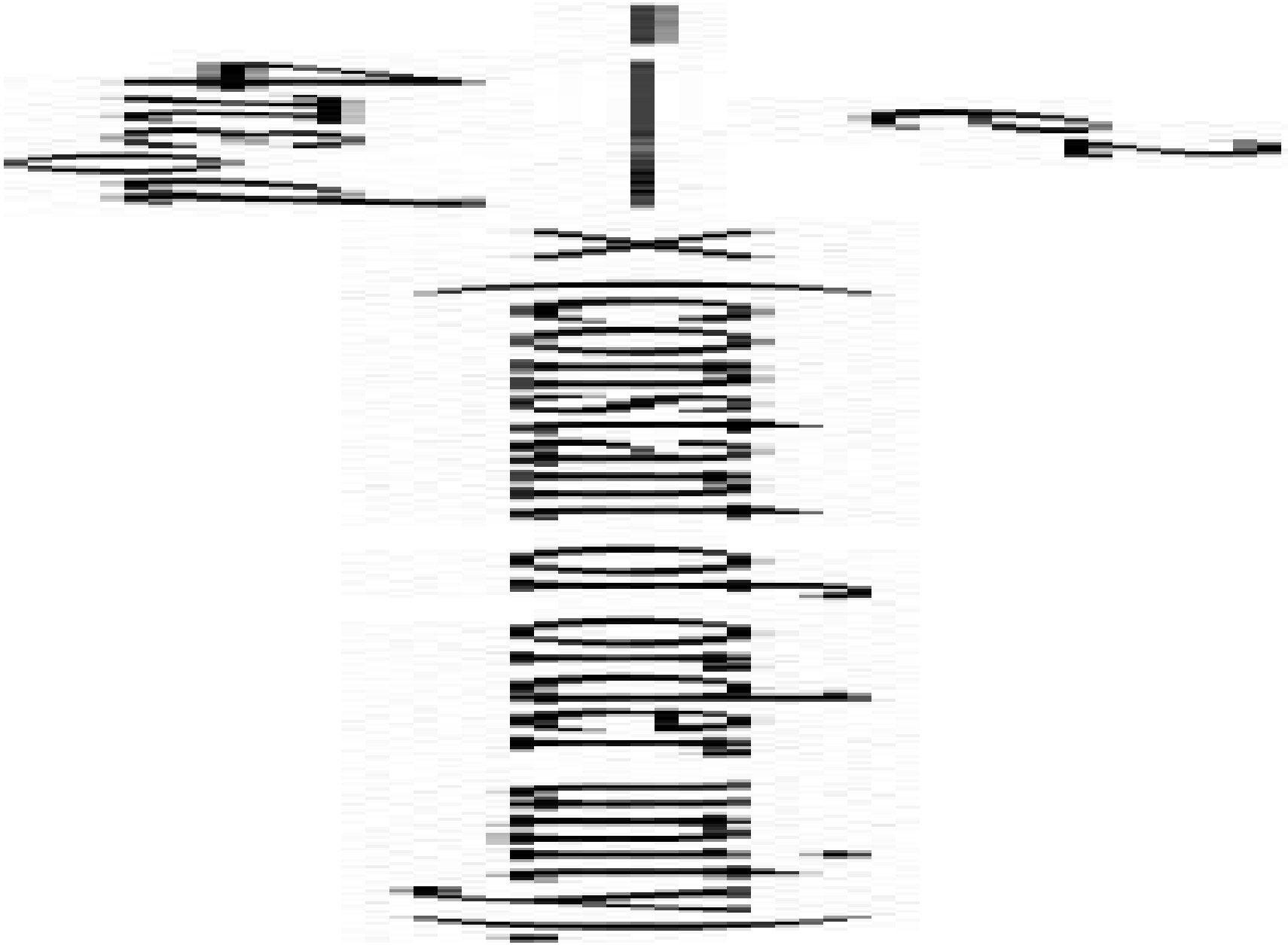


The restriction $j < i$ in the summation makes sure that the interaction between each pair of charges is only counted once. We can express the potential energy in a different way by deliberately counting each interaction twice, and then dividing the answer by two, writing

(1.36)

since $(1/4\pi\varepsilon 0)\sum j\neq i(qj/rji)$ is just the potential ϕi at the charge qi due to all the other charges qj.

In the sodium chloride crystal, all the qi are either +e or −e. If the separation between neighbouring sodium and chlorine ions is d, then the potential energy of each neighbouring pair is $-e^2/4\pi\varepsilon 0d$. Because there are exactly equal numbers of positive and negative ions surrounding a particular ion pair, there is a lot of cancellation in the summation 1.36, and the total potential energy per ion pair becomes



For sodium chloride, d = 2.8 × 10−10m, and the constant has the value 1.75, leading to a potential energy of −8.0eV per ion pair. Adding the 1.5eV needed to transfer an electron from the sodium atom to the chlorine atom, the heat of formation of sodium chloride is found to be 6.5eV per ion pair, or equivalently, 6.8×10^5 joules per gram molecule. This is within 10% of the experimentally measured value, and the magnitude of a few electron volts is typical of chemical energy changes*.

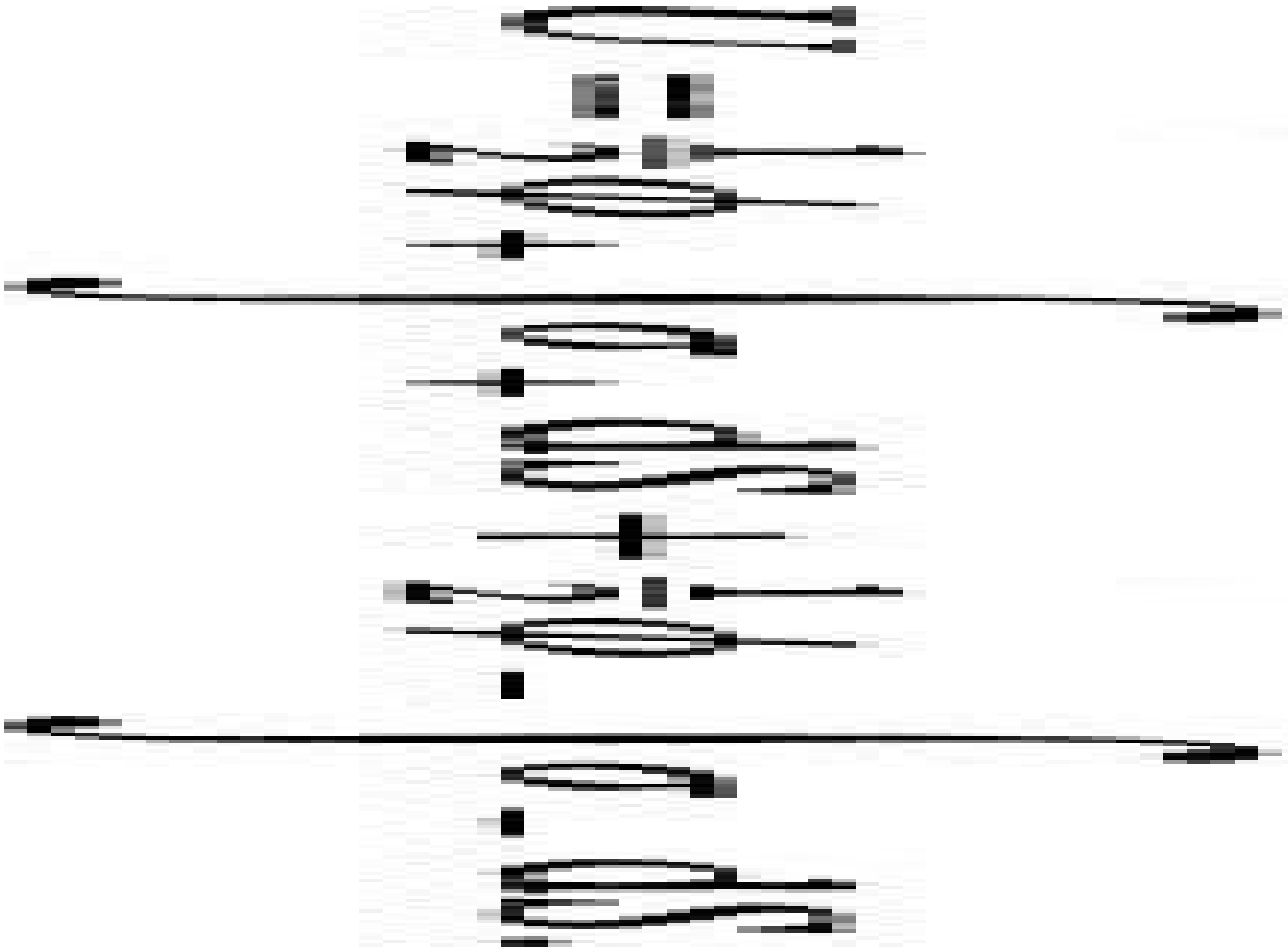
1.5.5 Capacitors, and energy in macroscopic fields

The potential energy in Equation (1.36) depends on the product of charges, and therefore if two distributions of charge are superimposed the total potential energy is not the sum of the potential energies of the two systems separately. This non-linearity means that the potential energy associated with a macroscopic distribution of charge cannot usually be found from the macroscopic charge density and the macroscopic potential, which are simply averages over the corresponding atomic quantities. However, we can at once work out the potential energy of macroscopic charges on conductors in vacuum. Since the macroscopic field is always zero inside a conductor, any energy changes occur because of the movement of charges outside, where there is no distinction between the atomic and the macroscopic field. The only charges are on the surfaces of the conductors, and the summation in Equation (1.36) can be replaced by an integration over surface charge density:

(1.37)

In electrical circuits, charges are stored on a capacitor, which consists of a pair of conductors. When the fields some distance away from the capacitor are negligible, as is usually the case, Gauss' theorem tells us that the total charge

carried by the capacitor is zero, and that the two conductors must carry equal and opposite charges. There is no component of electric field lying along the surface of a conductor, since the surface is an equipotential. Suppose the two conductors in a capacitor carry charges ±Q. These charges are distributed with surface charge densities $\sigma+$ and $\sigma-$, which may not be constant over the two conductors. However, the potentials of the conductors are constant, and have values $\phi+$ and $\phi-$ say. Substituting in Equation (1.37), the potential energy becomes



The surface integrals yield simply the total charges ±Q on the two conductors, and

(1.38)

We have introduced the symbol V for the potential difference between the two conductors, to distinguish it from the scalar field ϕ, which is a function of position.

The expression for the potential energy of the capacitor can be simplified even further. In a given capacitor, Q and V are always proportional to one another, because there is only one way in which surface charge can be distributed so that the macroscopic field is zero everywhere inside the conductors. The pattern of the lines of force between the conductors is determined only by their geometry, and the principle of superposition then ensures that charge and voltage are proportional; if the charge on each conductor is doubled, then the field retains the same pattern, but is everywhere doubled in strength. For any capacitor, we can write

(1.39)

The constant of proportionality C, which depends only on the geometry of the conductors, is called the capacitance. Now the energy stored by the capacitor can be expressed in terms of the constant C and only one of the variables Q or V

(1.40)

A simple arrangement of conductors, which has an easily calculated capacitance, is the parallel plate capacitor. Two plates, each of area A, are separated by a distance d. When there is a potential difference V between the plates, field lines pass from one to the other as shown in Figure 1.23. Except at the edges, the magnitude of the field is V/d. Provided that d is small compared with the dimensions of the plates, edge effects are unimportant, and we shall ignore them, assuming the field to be uniform between the plates, and zero elsewhere. It was shown in section 1.4.2 that the field at the surface of a conductor is related to the surface charge density σ by $\sigma = \varepsilon_0 E$, and it follows that there is a uniform charge density $\pm Q/A$ on the inner surface of the plates. Hence

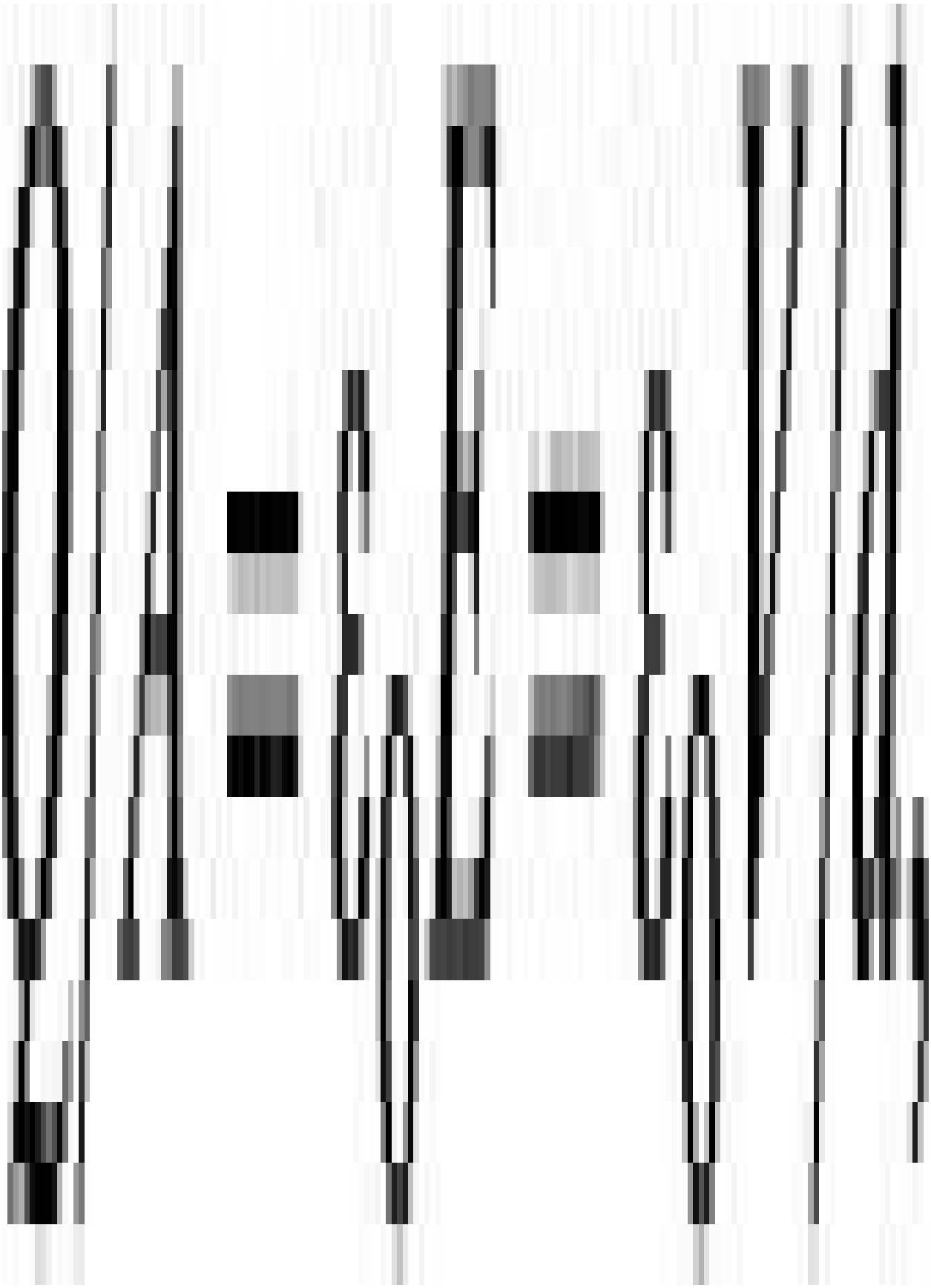
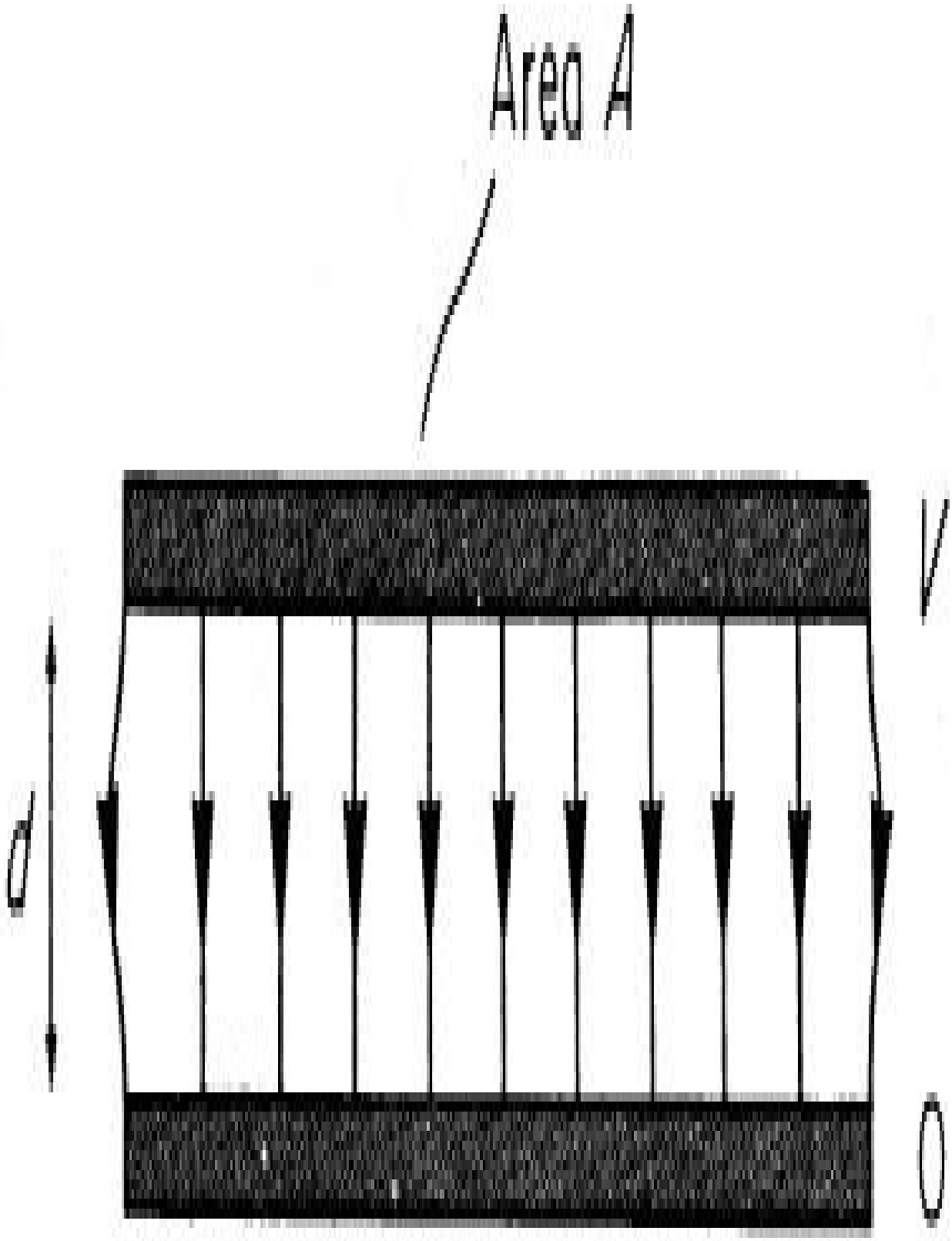


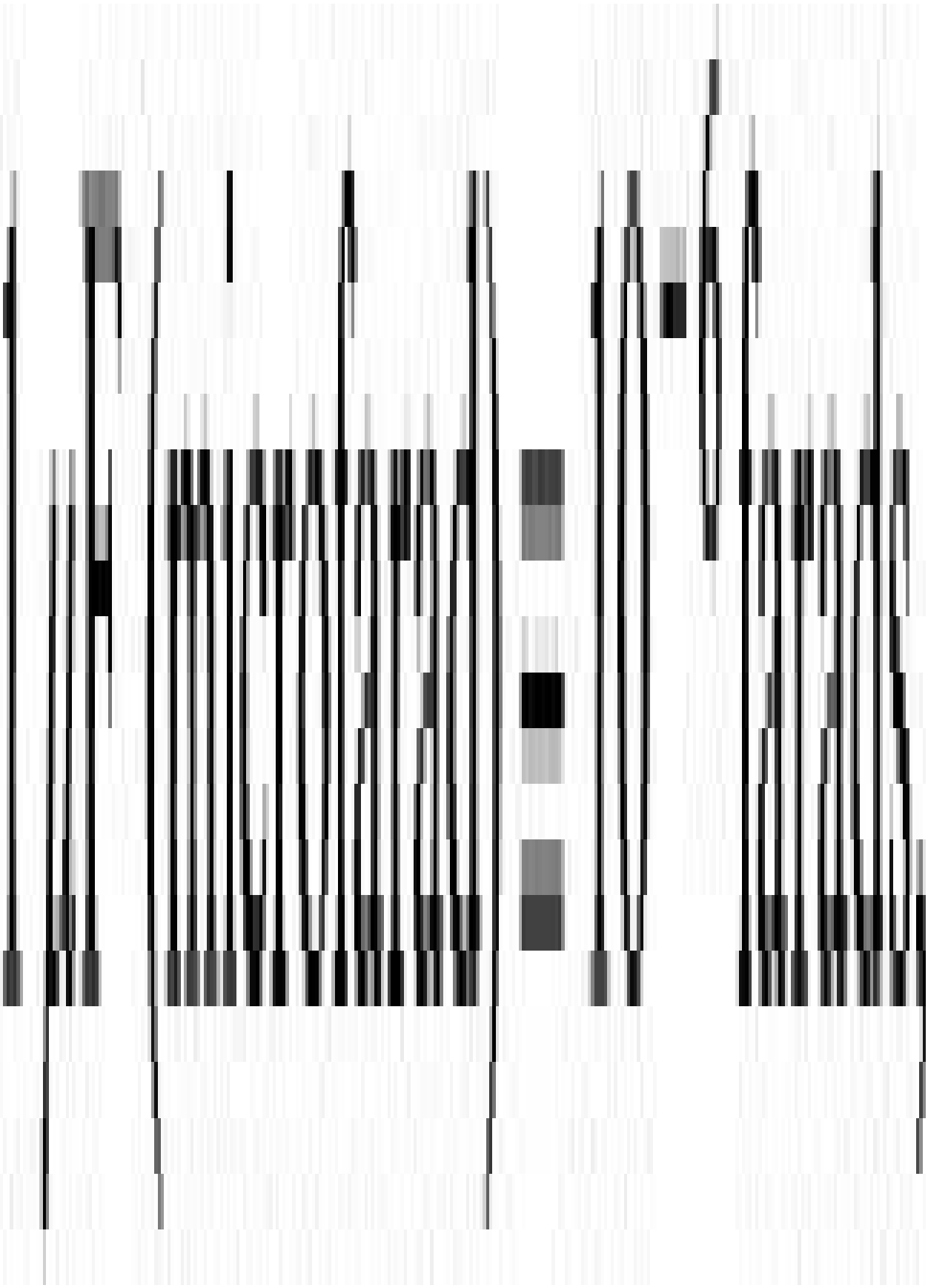
Figure 1.23. A cross-section through a parallel plate capacitor.

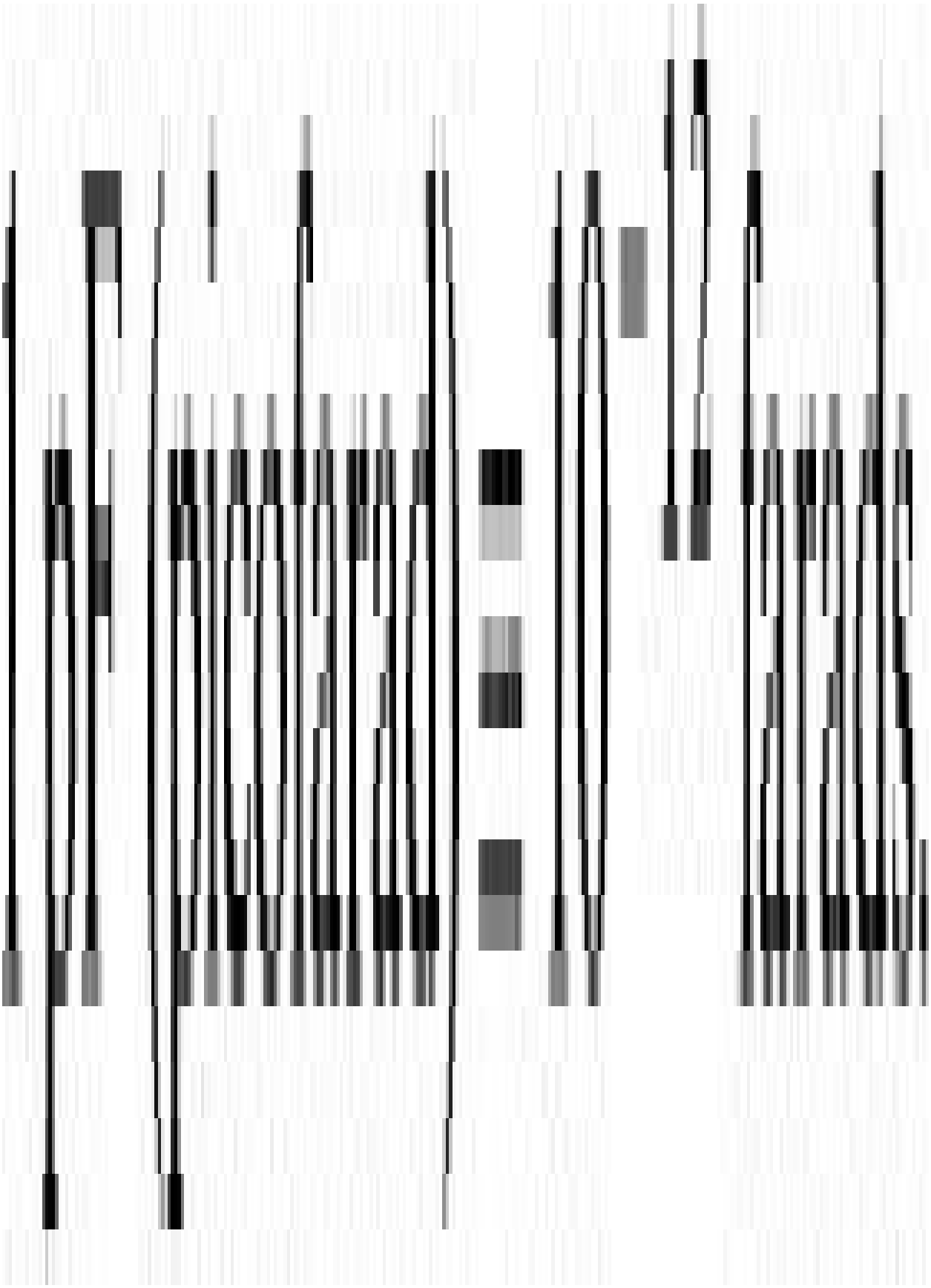

Area A
d
V
Q

and the capacitance

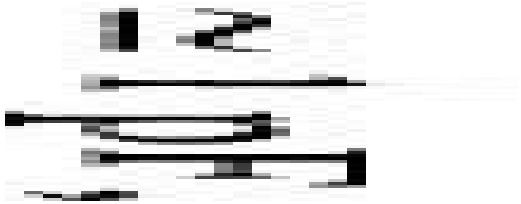
(1.41)

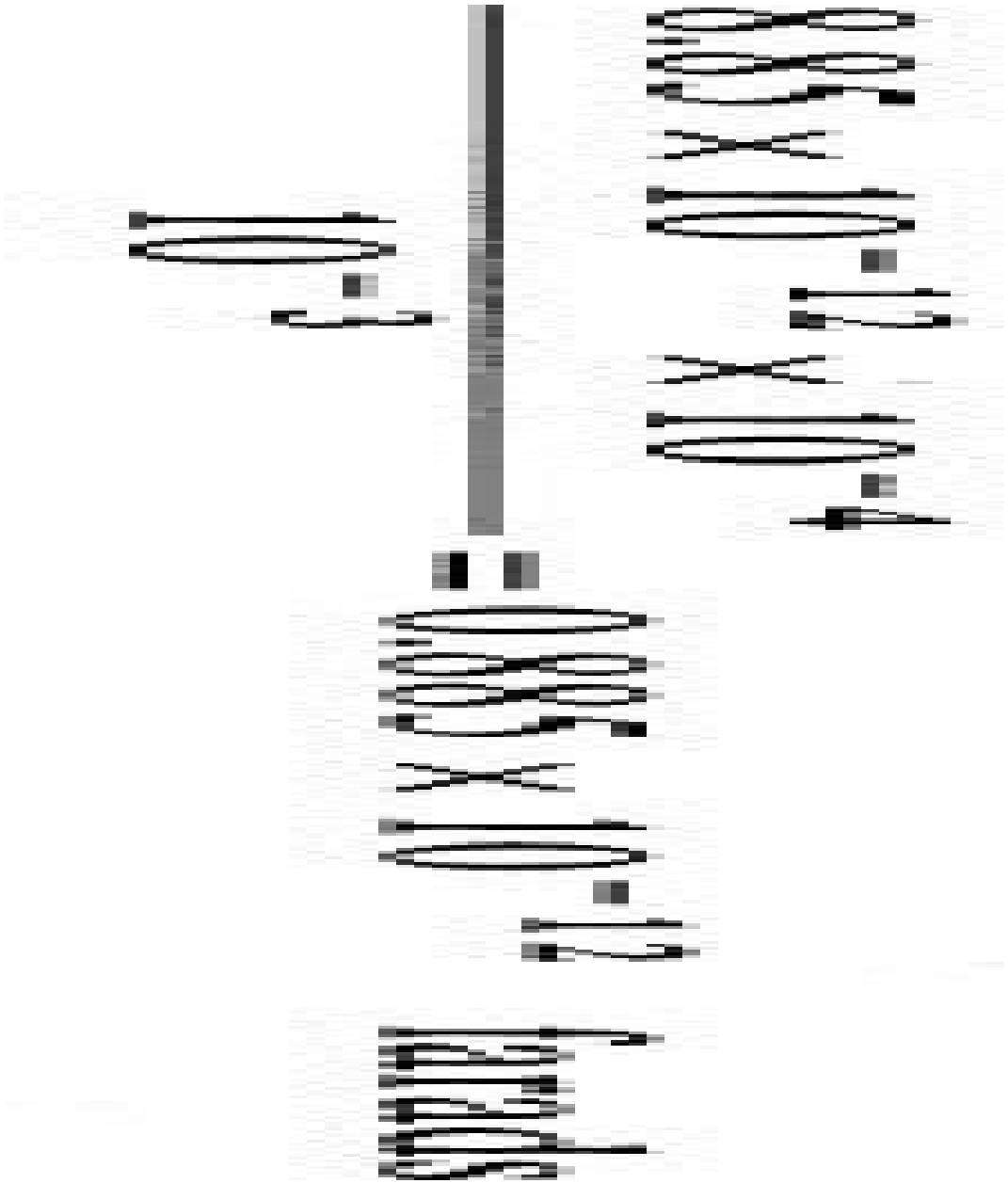
The unit of capacitance is the farad. The farad is a very large unit, and capacitances are often quoted in microfarads or picofarads:





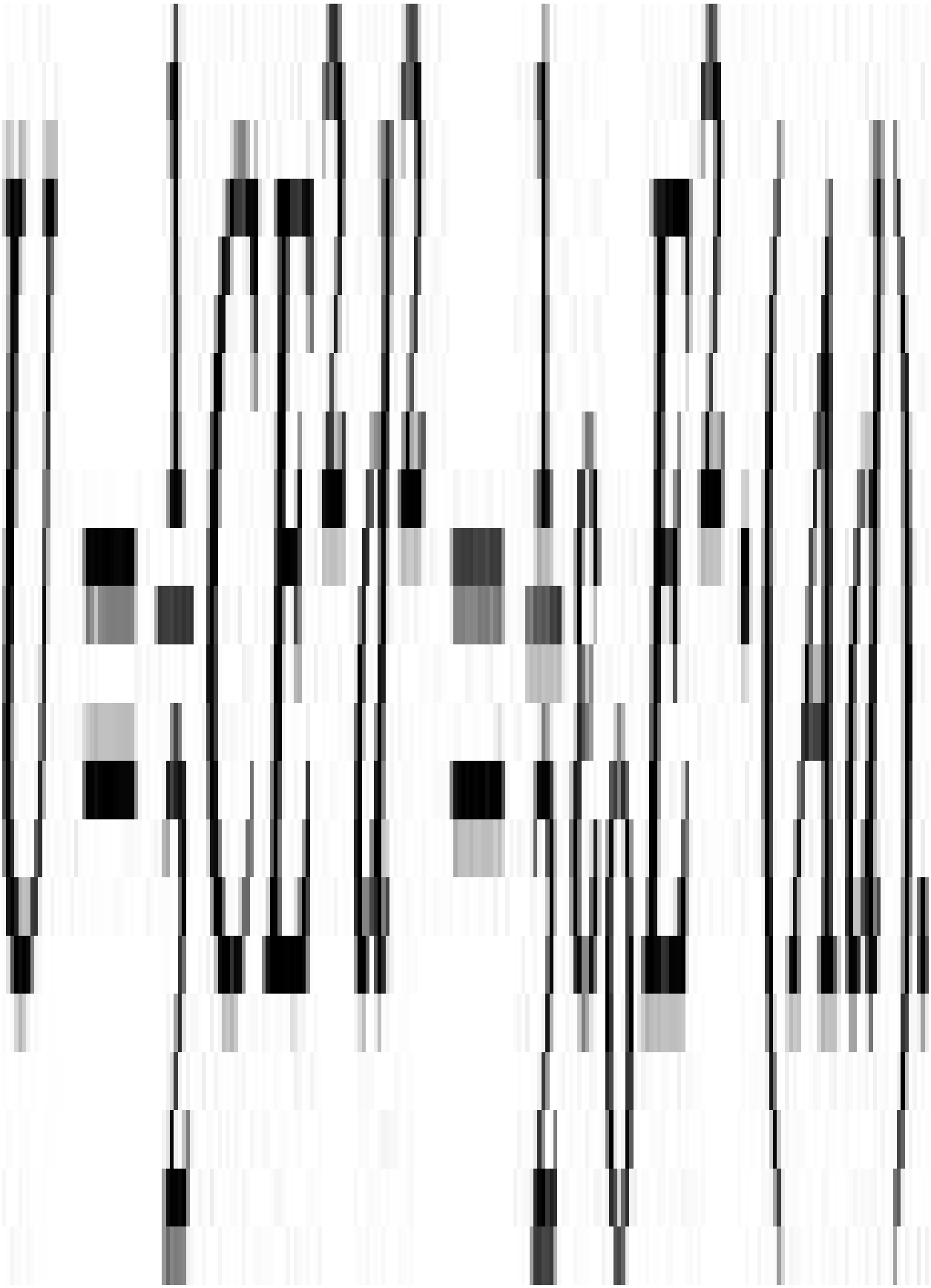
Practical capacitors are usually made with a solid material between the plates; the effect of inserting the solid is considered in Chapter 2. However, in an electrical circuit there is a capacitance between any pair of conductors. Such a capacitance is usually unwanted and is then called a stray capacitance. Stray capacitances may have a deleterious effect on performance, so that it is important to have a rough idea of their magnitude. Of course circuits are usually in air rather than vacuum, but this makes practically no difference to the magnitudes of the stray capacitances. As an example, let us calculate the capacitance of two plates, each of area 1cm^2 and separated by 1mm. According to Equation (1.41), the capacitance is





and the energy stored when 100 volts is applied across the plates is $U = CV^2 = 5 \times 10{-}9$ joules.

For the parallel plate capacitor, $V = Ed$, and, expressing the energy in terms of the electric field instead of the voltage, we find



The volume of the region between the plates is (Ad) and the amount of work needed to build up the field E can be written as an integral over this volume:

(1.42)

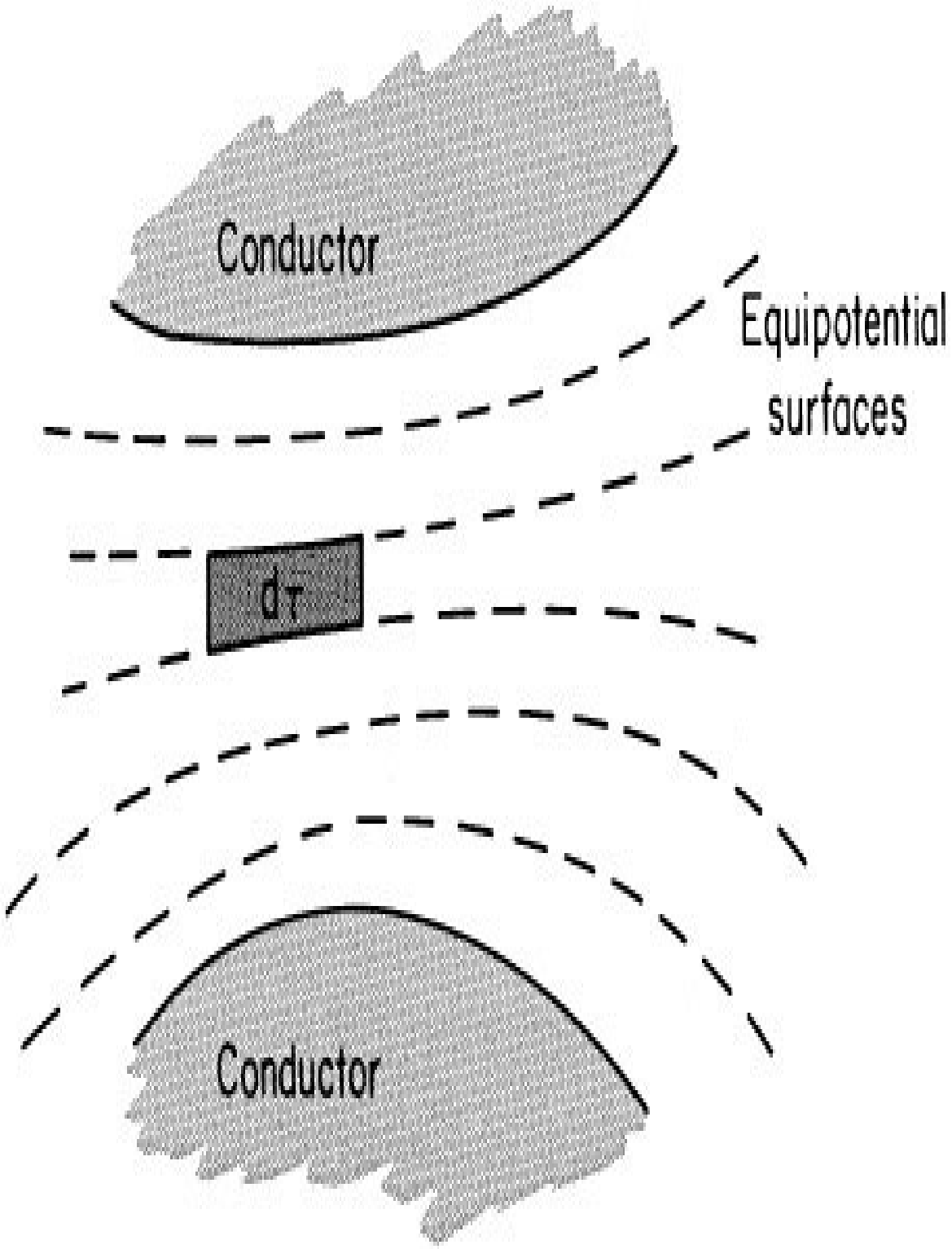
Although we have derived this result by considering the field in a parallel plate capacitor, it is true for any arrangement of conductors in vacuum, provided that the integration extends over the whole of the region where the field is non-zero. A rigorous proof of Equation (1.42) is given at the end of this section, but one can see that it should be true for any arrangement of conductors by imagining very thin conducting sheets placed along the equipotential surfaces of the field. Induced charges appear on the conducting sheets, but the pattern of equipotentials, and hence of field lines, is unaltered. If the equipotential surfaces are chosen close enough together, as in Figure 1.24, each volume element $d\tau$ looks like a parallel plate capacitor, and has potential energy $\varepsilon_0 E^2 d\tau$. Integration then gives the total energy stored by the real capacitor.

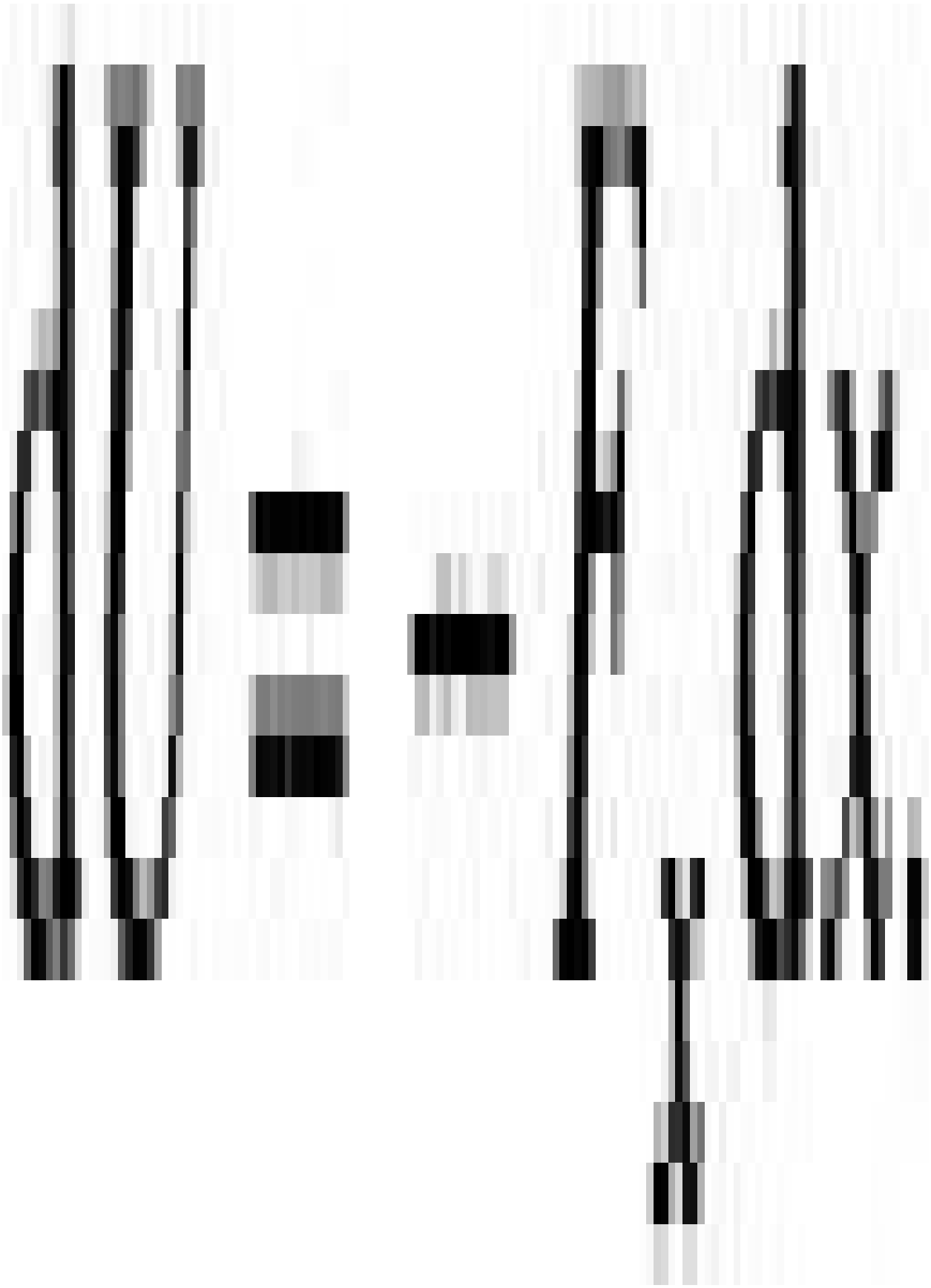
It is sometimes helpful to think of the energy $\varepsilon_0 E^2 d\tau$ as residing in the volume $d\tau$, that is, to associate an energy density $\varepsilon_0 E^2$ with the field. At first this may seem a strange idea, since the whole concept of electric field is an abstract one. However, the energy density has been defined so that it must give the correct value for the total energy, and it is a useful concept, because it has eliminated reference to individual charges.

In air, electric fields are limited to about 10^6 volts m^{-1}—above this value electrical breakdown will occur. The corresponding energy density is $\varepsilon_0 \times 10^{12}$, or roughly 4 joules m^{-3}. This is a very small amount compared with the electrostatic energy liberated in fission (2×10^{13} joules from 250 g of uranium) or in chemical reactions (6×10^5 joules for the heat of formation of a gram mol of salt—about 60 g).

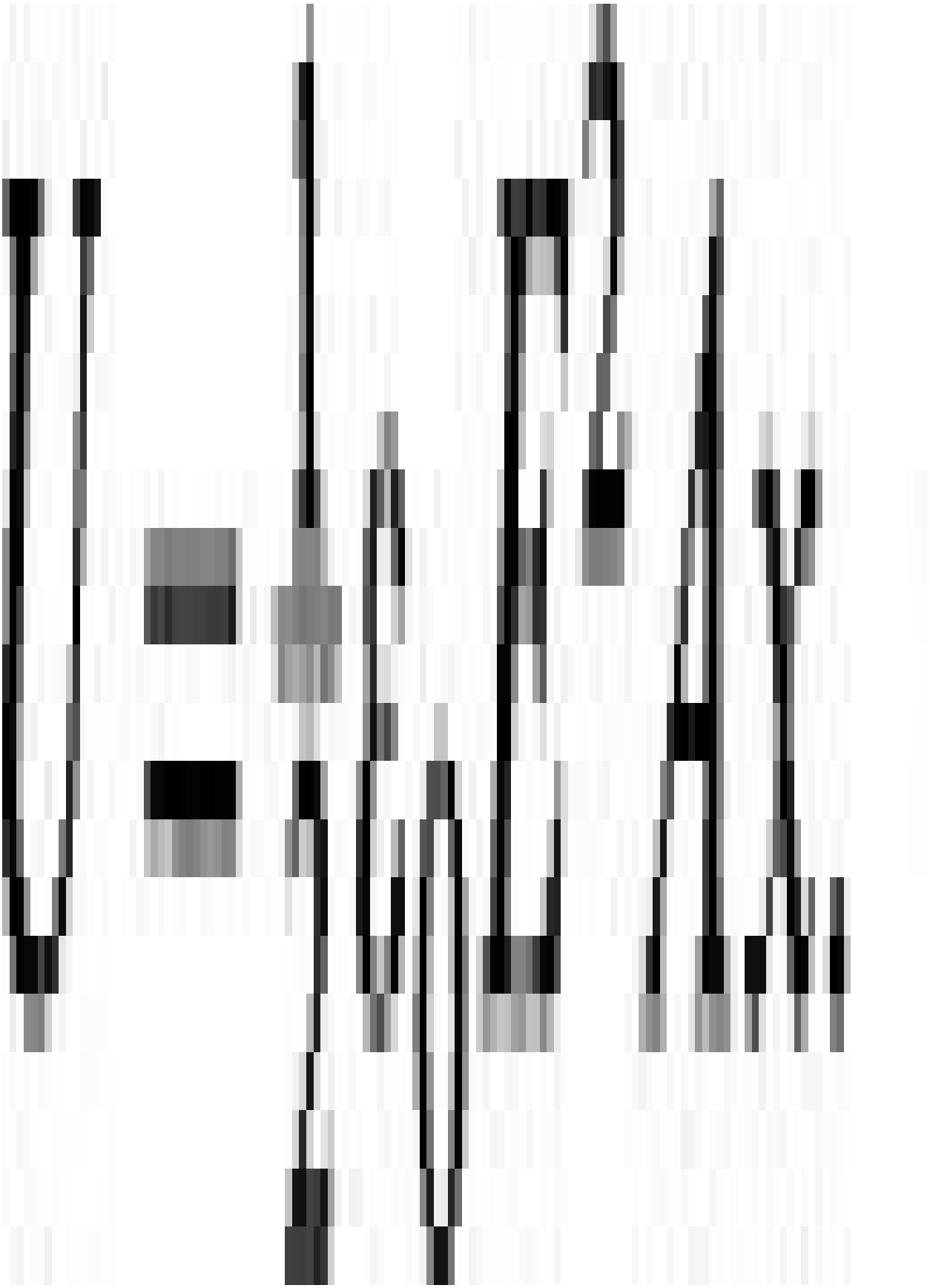
Because the two plates of a parallel plate capacitor are oppositely charged, they are attracted together. If the separation of the plates is x, the force Fx on each plate is in the direction of decreasing x, and the mechanical work expended in pulling them apart a further distance dx is −Fxdx. For an isolated capacitor, the mechanical work must equal the change dU in stored energy, i.e.

Figure 1.24. The space between conductors can be imagined as being divided into a lot of little parallel plate capacitors like the one occupying the volume δτ.


Conductor
Equipotential
surfaces
dτ
Conductor



For plates of area A and separation x, the volume in which there is an electric field is Ax. Within this volume the energy density is $\varepsilon_0 E^2$, and the total stored energy is



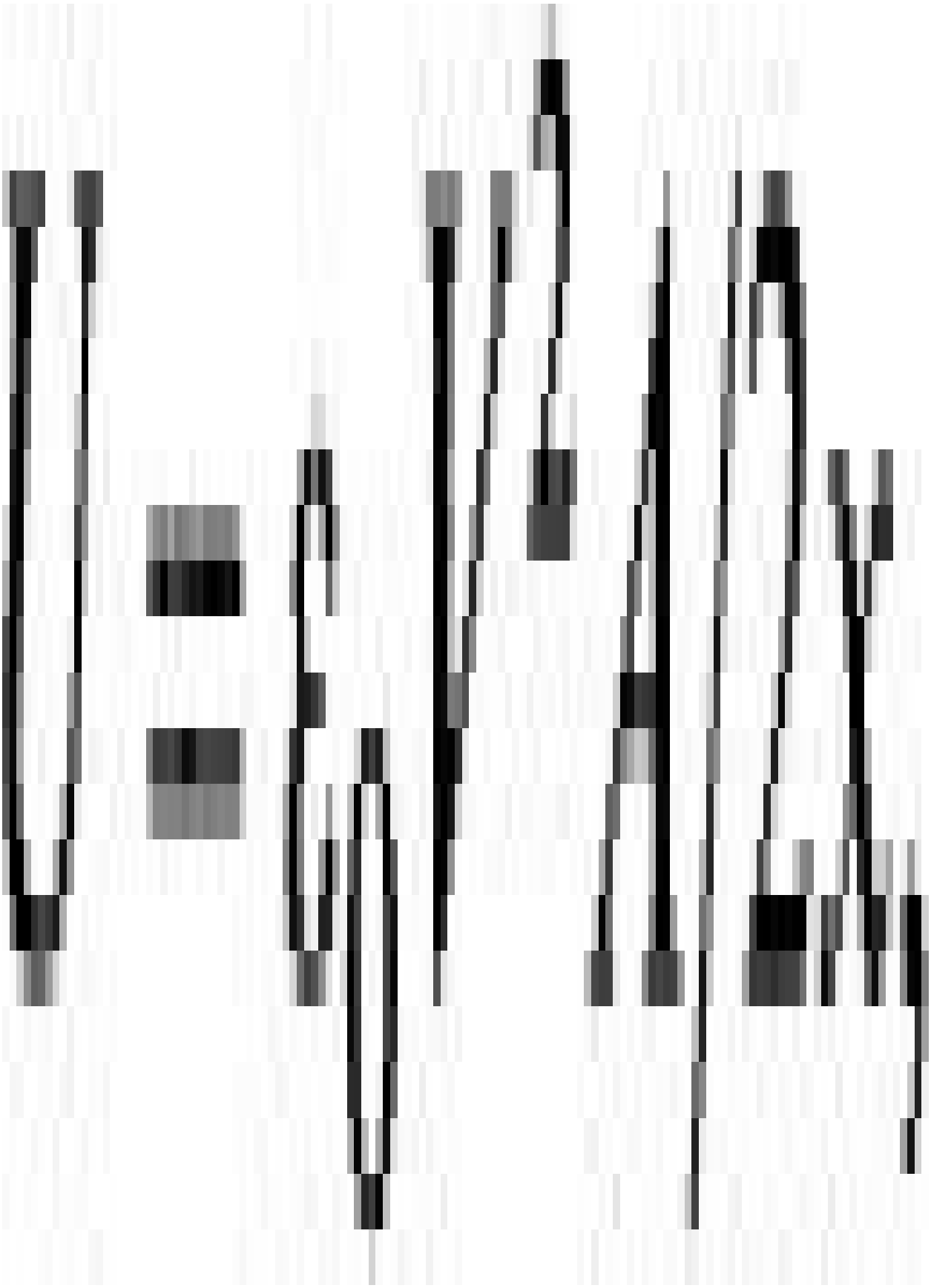
The field remains constant when the separation changes, since the charge density on the plates is $\sigma = \varepsilon_0 E$ (Equation (1.16)), and σ is fixed for an isolated capacitor. Thus

(1.43)

i.e. a force $\varepsilon_0 E^2$ per unit area attracting the plates together.

When the field at its surface is 5×10^6 volts m^{-1}, the force on a conductor is about 100 newtons m^{-2}, i.e. about 10kg wt. per square metre. The charge density on the conductor is $\sigma = \varepsilon_0 E$, and we can rewrite the force per unit area as σE. At first sight the occurrence of the factor is surprising, since we envisage the charge as located right on the conducting surface, where the field is E. However, on the atomic scale, the surface charge is actually spread through an extremely thin layer. The electric field changes throughout the layer in such a way that the net force per unit area is indeed σE (the proof of this is left as an exercise in the examples at the end of the chapter).

Equation (1.43) is only valid when the electric field remains constant as the separation of the capacitor plates changes. If the capacitor is connected to a battery so that its plates are held at a constant potential difference V, say, the field is V/x, the stored energy becomes



and

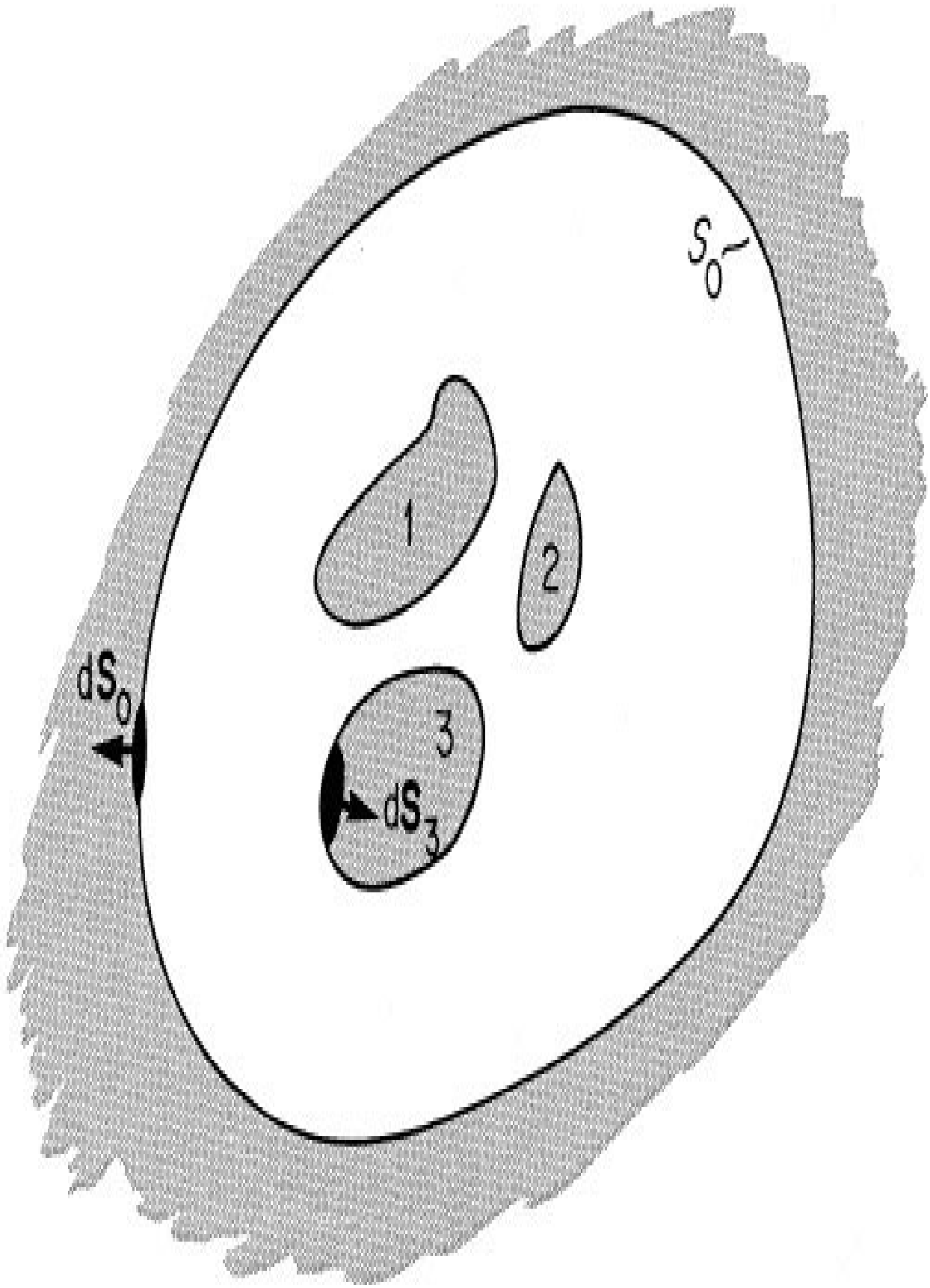
(1.44)

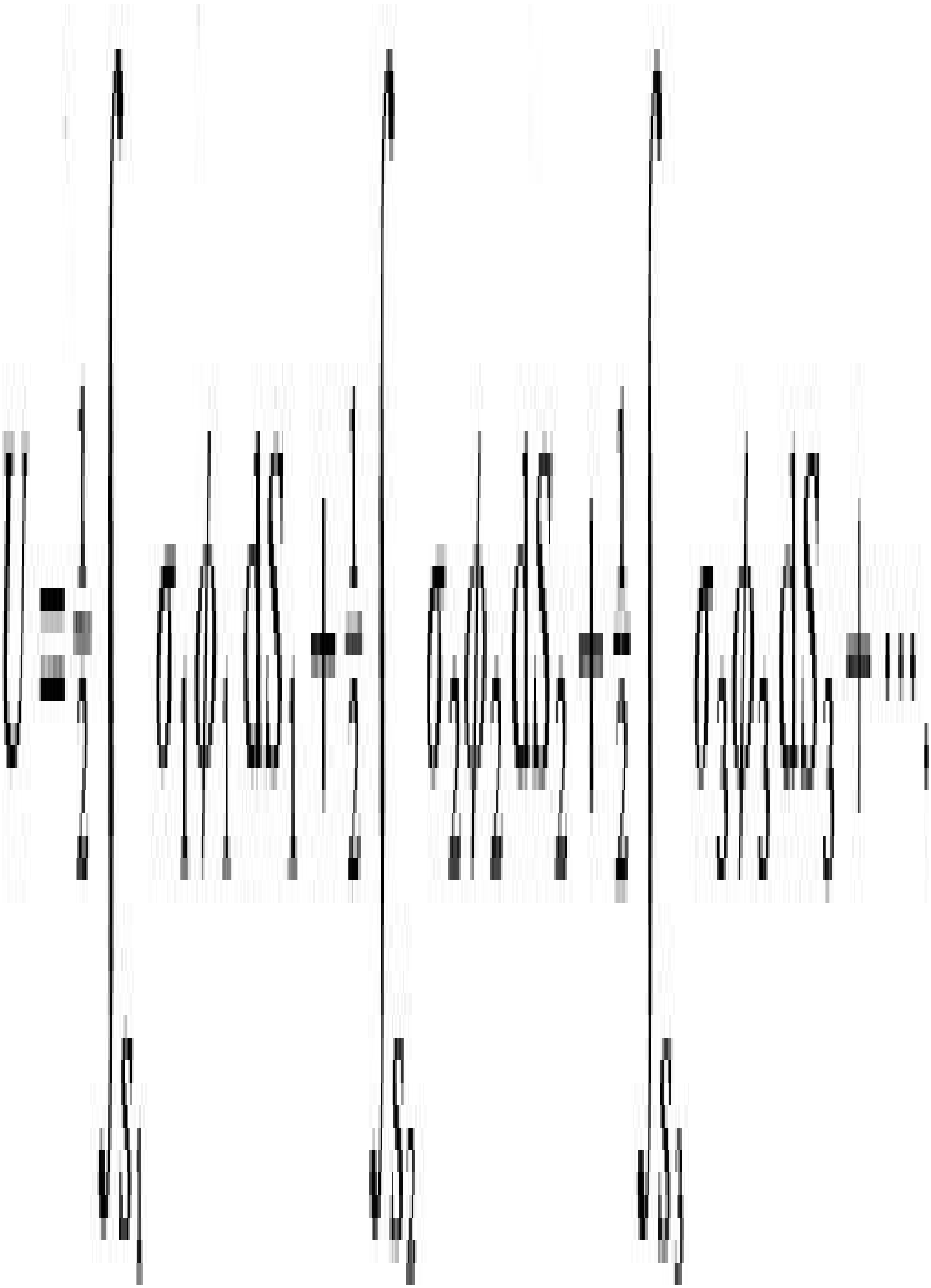
The difference between Equations (1.43) and (1.44) arises because of work done by the battery or energy returned to it when the capacitor plates move.

1.5.6 Energy stored by a number of charged conductors

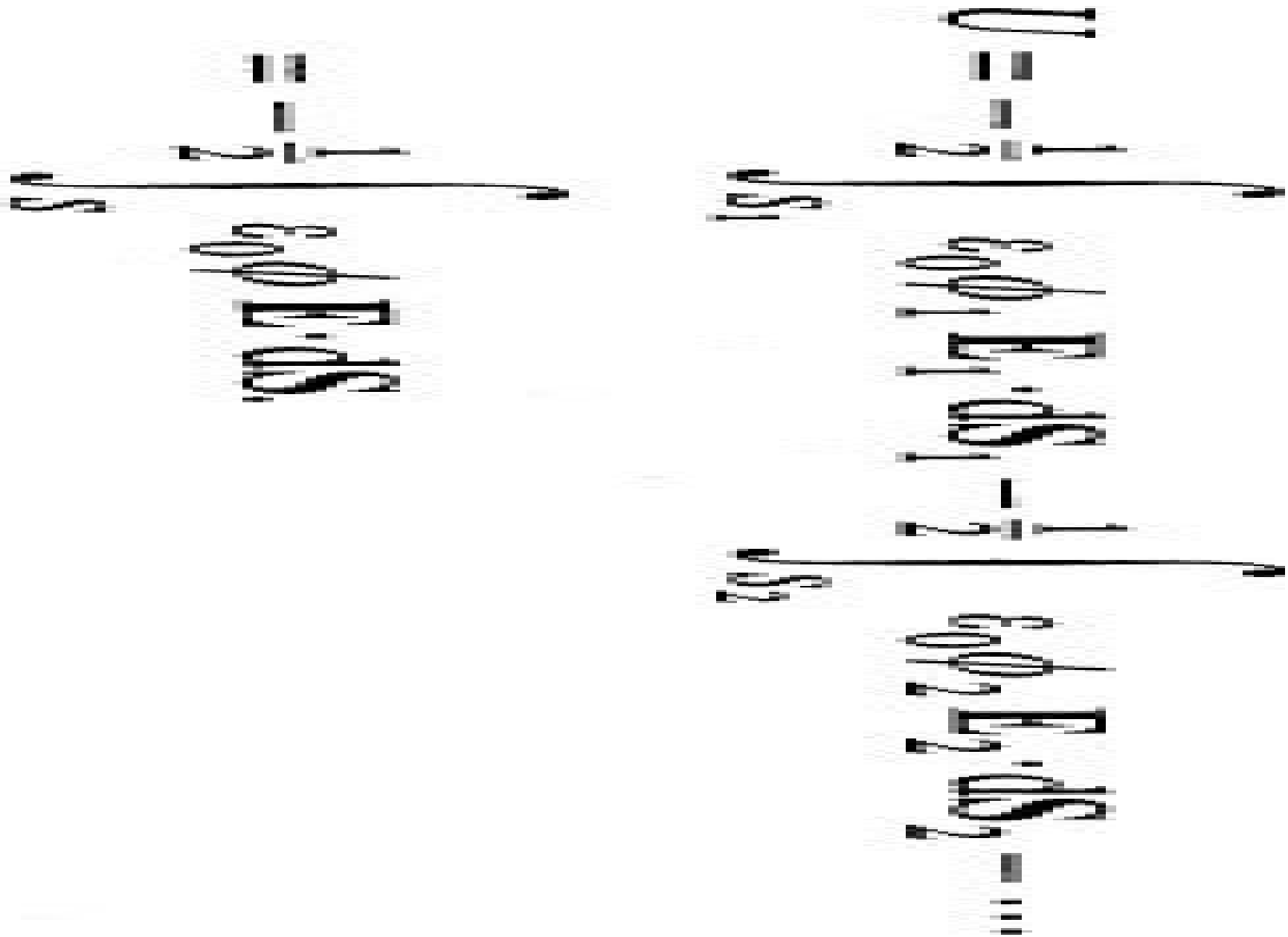
The rigorous proof of Equation (1.42) requires some manipulation with vector operators. Consider a number of conductors 1,2,3… at potentials $\phi 1$, $\phi 2$, $\phi 3$…, and carrying charges on their surfaces of densities $\sigma 1$, $\sigma 2$, $\sigma 3$…. Such an assembly of conductors is illustrated in Figure 1.25. In general the charge densities will not be constant over each surface. The total potential energy of the system of conductor is

Figure 1.25. An assembly of conductors.

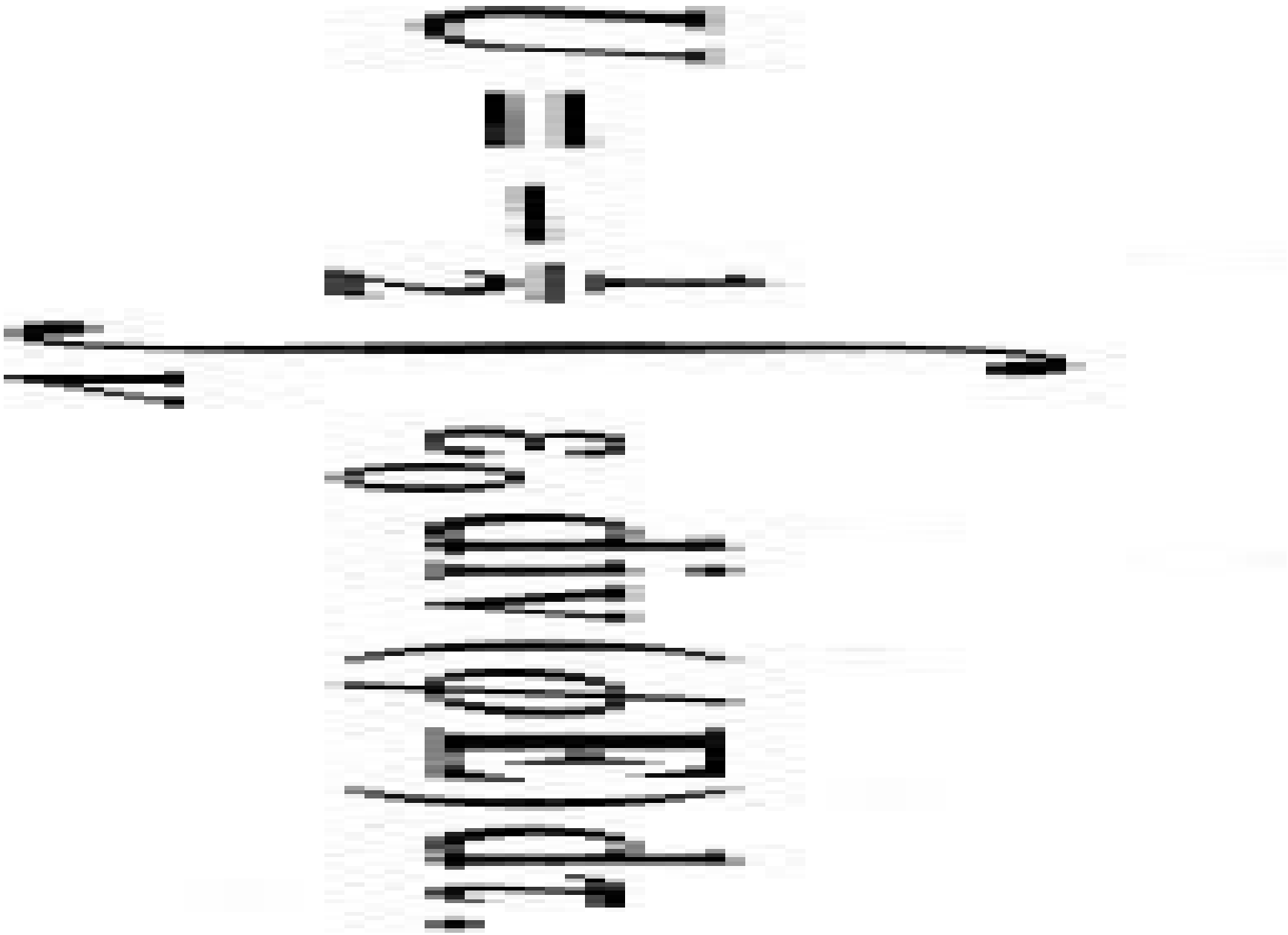

S_0
1
2
3
dS_0
dS_3



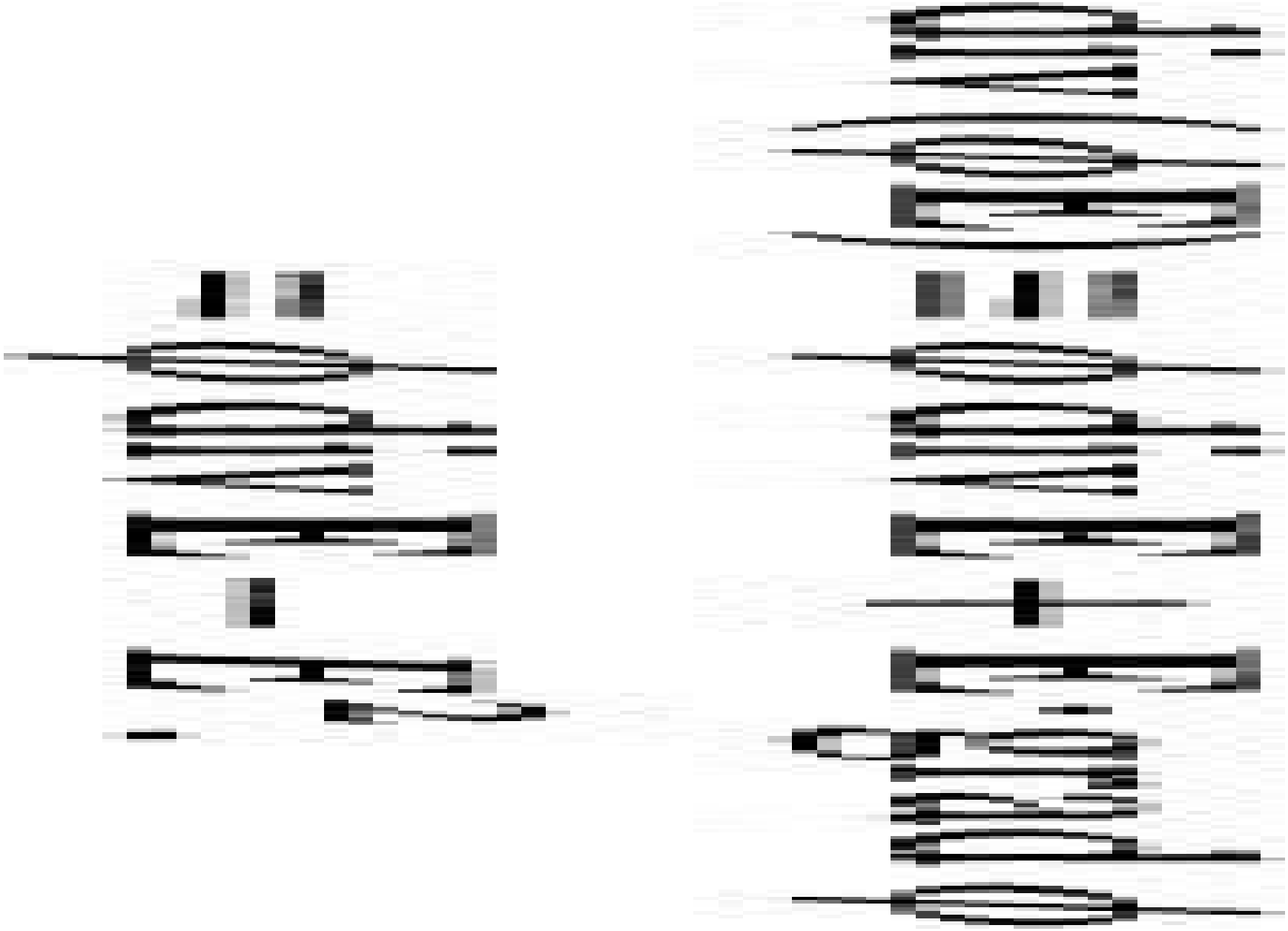
The conductors are surrounded by another surface S0, sufficiently far away that the field is negligible on S0. The set of surface integrals for each conductor can be replaced by a single integral over the surface $S = S0 + S1 + S2 + S3\ldots$ bounding the volume V where the field is non-zero. The outward normals to V are directed towards the inside of each conductor, and since the charge density σ is related to the outward field at the surface by $\sigma = \varepsilon 0E$, the potential energy can be written



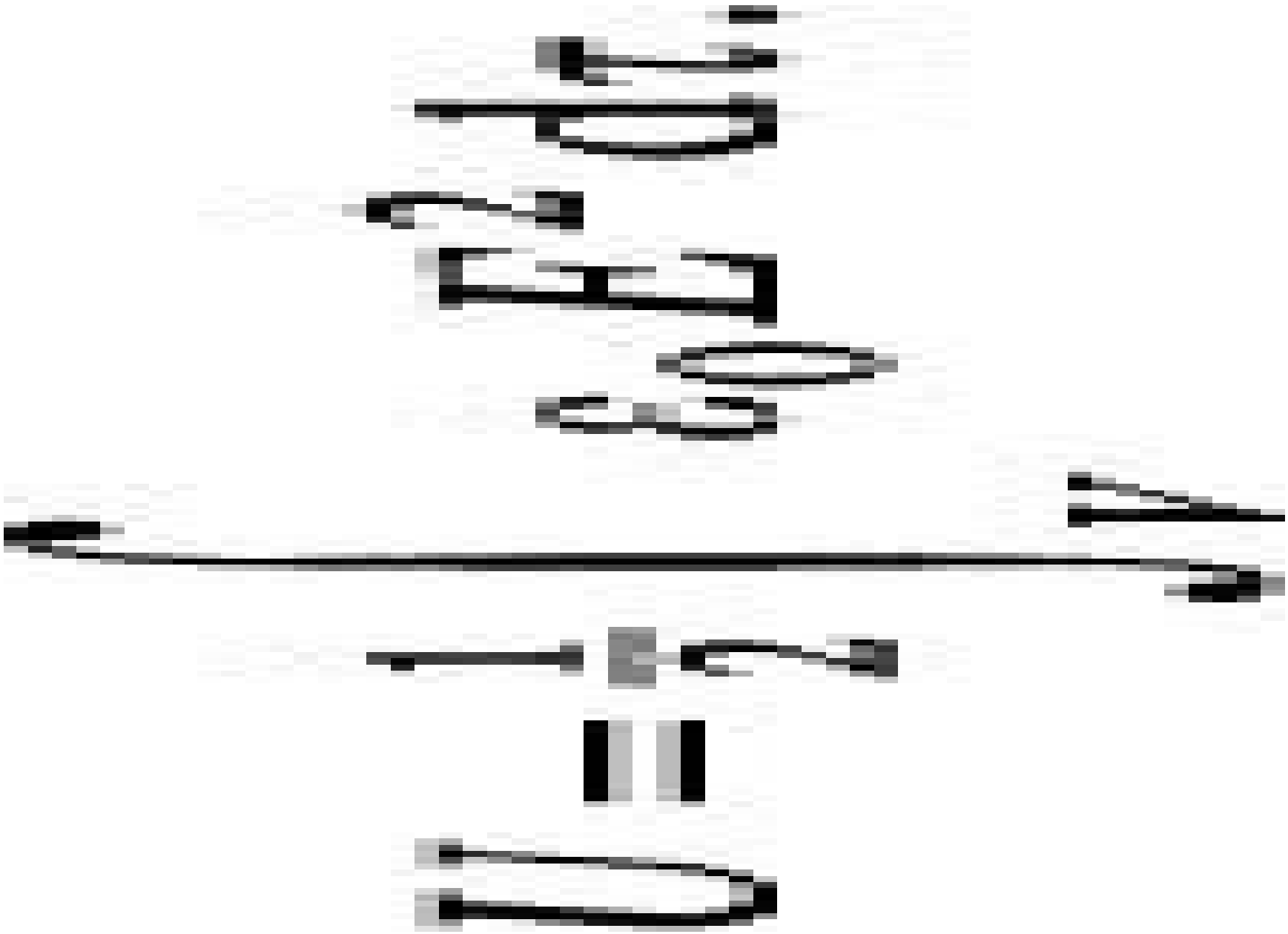
Using the divergence theorem, the surface integral is converted into a volume integral:



We now use the vector identity



The conductors are in vacuum, so there are no charges within V, and div E = 0. In the volume integral, div(ϕE) can therefore be replaced by ($-E^2$), and finally



Notice that the proof holds for any distribution of charges on the conductors, and does not require that the net charge be zero, as it is in the parallel plate capacitor. When there is a net charge on the surfaces S1, S2, S3…, an equal and opposite charge is always induced on S0. But by choosing S0 to be sufficiently far away, the contribution of this induced charge can be made indefinitely small.

PROBLEMS 1

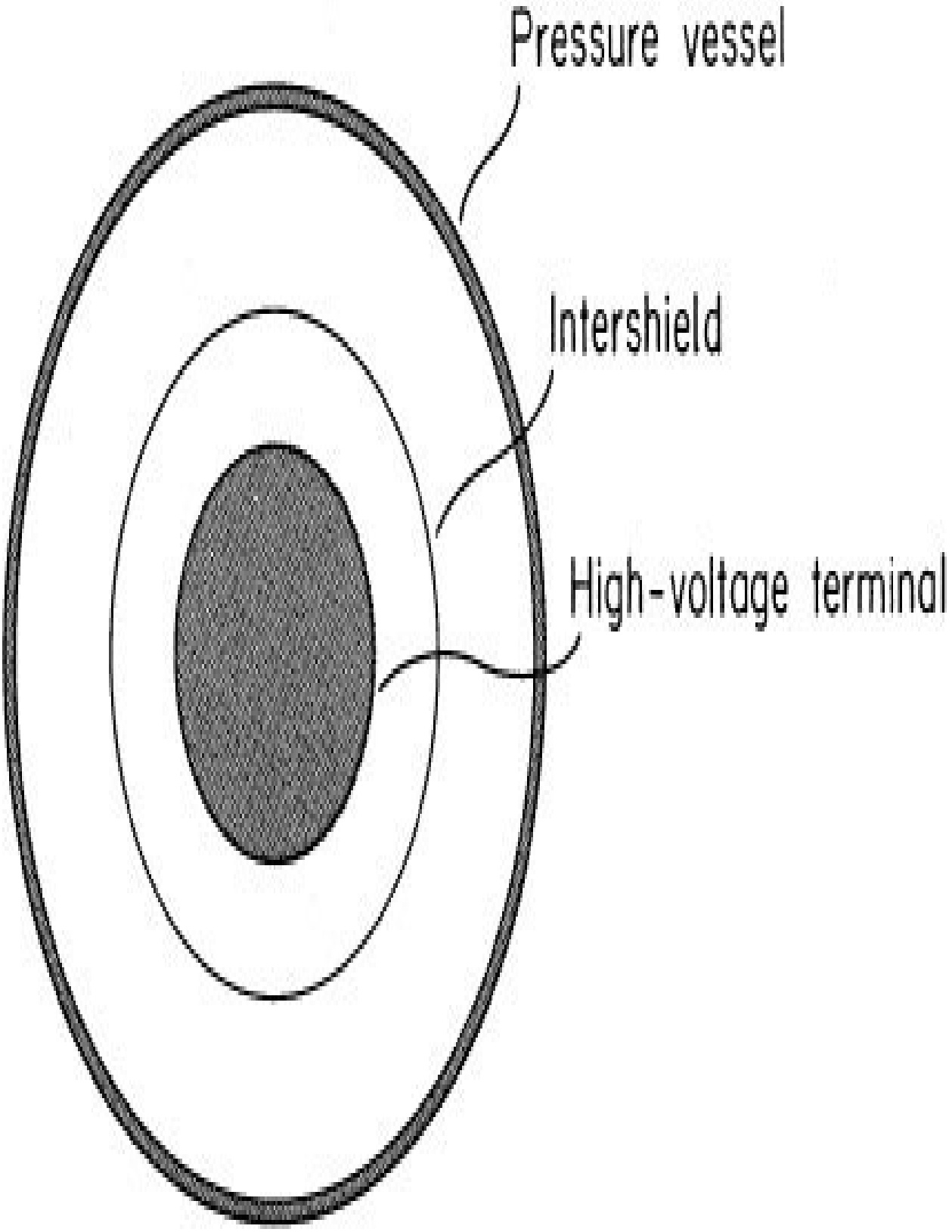
1.1 Three point charges are placed at the vertices of an equilateral triangle with sides one metre long. The charges are all positive and each of magnitude 1 μC. What is the force on each charge? What is the magnitude of the electric field (a) at the centre of the triangle, (b) at the mid-point of one side of the triangle?

1.2 Draw schematically the equipotentials and field lines around two parallel wires each carrying line charges λ coulomb/m (a) if the charges are of the same sign, (b) if they are of opposite sign.

1.3 The maximum electric field which can be supported by dry air at atmospheric pressure is about 10^6 volts/m. What is the maximum potential difference to earth for a conducting sphere of radius 10cm in air? (Take the distance from earth to the sphere to be infinite: provided that the distance is large compared to 10cm, the correction is small.)

1.4 Figure 1.26 shows a cross-section of the cylindrical high-voltage terminal of a van de Graaff generator, surrounded by an ‘intershield’ and a pressure vessel, both of which are also cylindrical. The gas in the pressure vessel breaks down in electric fields greater than 1.6×10^7 volts/m. If the radii of the terminal, intershield and pressure vessel are 1.5m, 2.5m and 4m respectively, what is the highest potential difference that can be maintained between the terminal and the pressure vessel? (Hint: the intershield must be maintained at a potential such that breakdown is about to occur on its own outer surface as well as on the surface of the terminal.)

Figure 1.26.


Pressure vessel
Intershield
High-voltage terminal

1.5 Two concentric conducting spherical shells have radii a and b (a < b). Work out their capacitance.

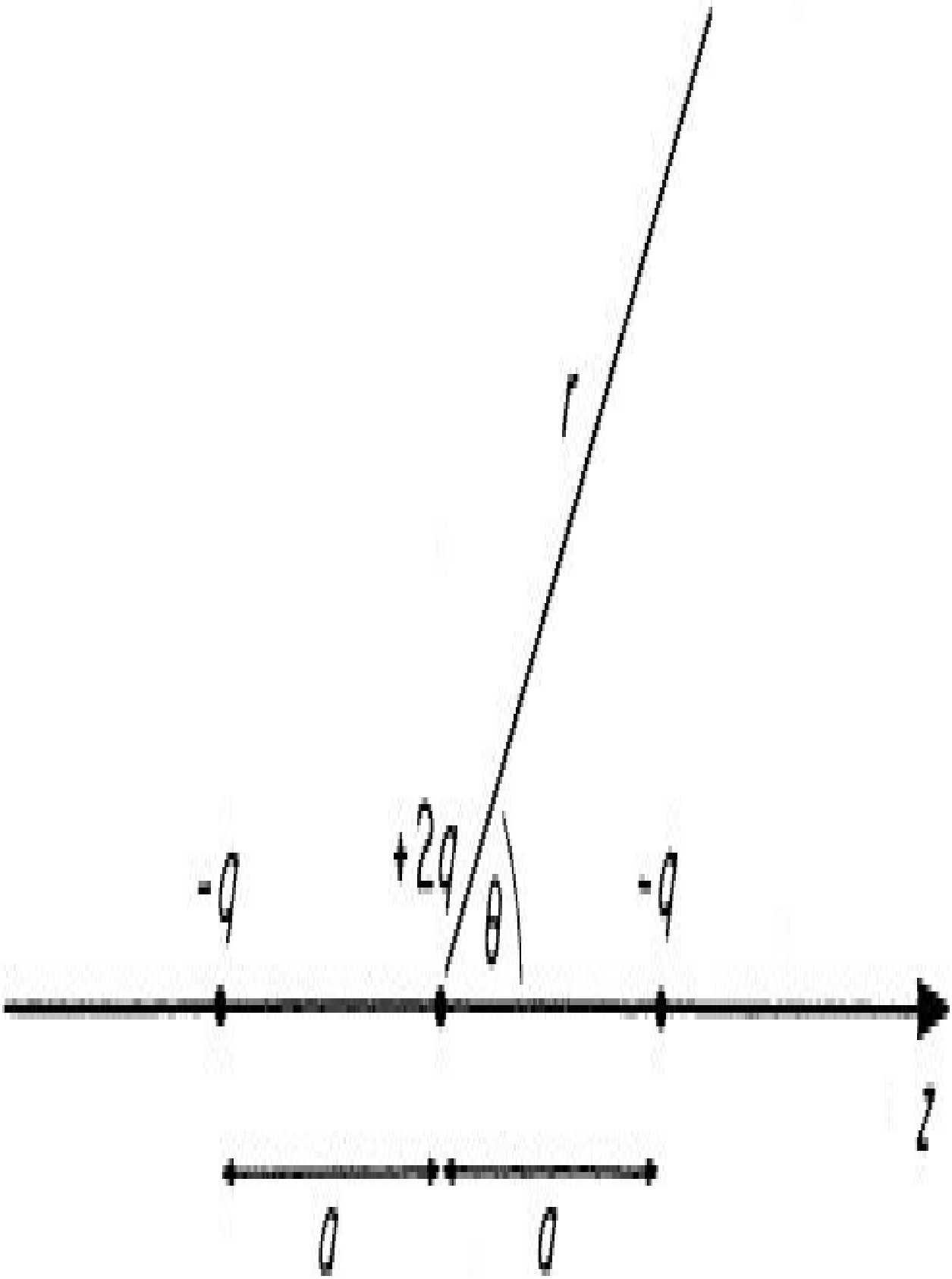
1.6 Estimate the capacitance of two conducting spheres each of radius 1cm whose centres are 10cm apart. Assume that the field between the spheres has the same shape as for equal and opposite point charges situated at their centres. This is a good approximation for spheres separated by a distance much greater than their radii—the charge distribution on one sphere is then hardly affected by the field due to the other.

1.7 The uranium nucleus contains 92 protons and has a radius of about 10^{-14}m. Assuming that the positive charge of the protons is uniformly distributed throughout the nucleus, calculate its electrostatic energy in MeV. (Imagine that the nucleus is built up layer by layer. When a sphere of radius r has already been built up, how much work is required to add an additional spherical shell of thickness dr?) If the uranium nucleus splits into two equal fragments each of radius 8×10^{-15}m, how much electrostatic energy is released?

1.8 Two large parallel plates of area A are a distance d apart and are maintained at potentials 0 and V. A third similar plate, carrying a charge q, is isolated from the other two and placed midway between them. What is the potential of this plate?

1.9 Three charges are placed on the z-axis as in Figure 1.27. This arrangement of charges, which is equivalent to two electric dipoles with dipole moment qa a distance a apart, is called an electric quadrupole. Calculate the potential at (r, θ, ϕ) for $r \gg a$.

Figure 1.27.

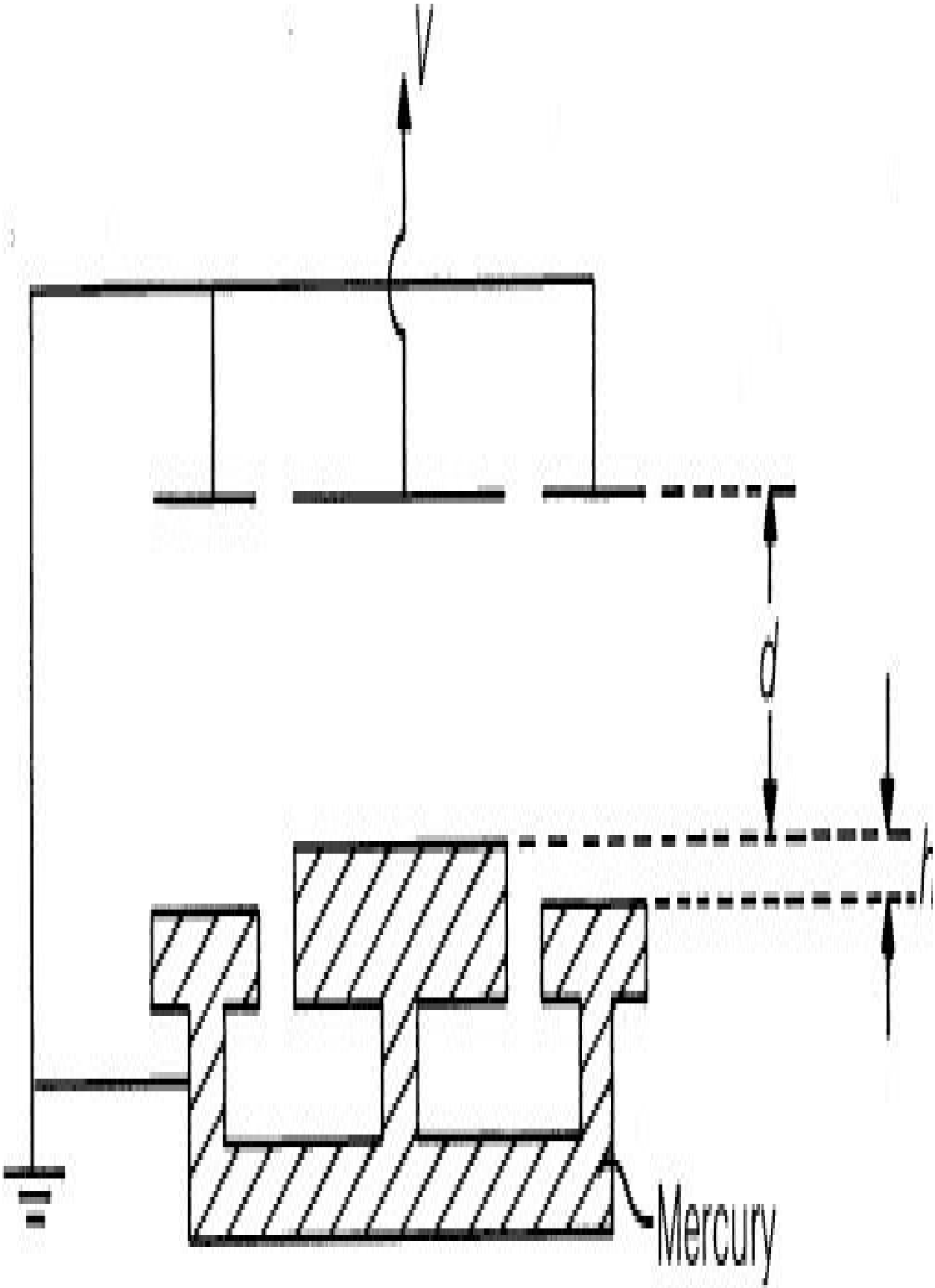

r
+2q
θ
-q
-q
z
a
a

1.10 An electrostatic voltmeter is illustrated in Figure 1.28. It is a capacitor with two horizontal plates, separated by a distance d; the lower plate is a pool of mercury. Outside the capacitor is a ‘guard ring’ which has both plates earthed when the voltage V to be measured is applied across the central region. The force between the plates causes a difference h in the level of the mercury in the two regions. Obtain an expression for the voltage V in terms of d, h, the density of mercury ρ and the acceleration of gravity .

1.11 The electric field at the surface of a conductor is E. The conductor carries a surface charge which is uniformly distributed through a layer of thickness d. Use Gauss’ theorem to find how the electric field varies in the surface layer, and hence show that the force on the conductor is $\varepsilon_0 E^2$ per unit area.

1.12 The field emission microscope consists of a fine hemispherical tungsten point at the centre of an evacuated sphere; the inside surface of the sphere forms a fluorescent screen. If the screen is maintained at a positive potential with respect to the point, there is a large electric field at the point, and when the magnitude of this field reaches about 10^8 V/m, electrons are drawn from the point even at room temperature; this process is called field emission. The electrons leaving the point follow the field lines, and a magnified image of objects on the point appears on the screen. In a typical microscope the radius of curvature of the point and the screen are 10nm and 10cm respectively. What is the magnification of the microscope? What is the minimum screen potential needed to cause field emission?

Figure 1.28.


V
d
Mercury

1.13 A parallel beam of electrons of kinetic energy 2keV is moving in a region of constant electrostatic potential and is incident on a metal grid which is held at a potential of 2kV. A second grid behind the first one is parallel to it and is at a potential V. If the electron beam direction is perpendicular to the grids, what is the smallest potential V required to prevent the beam from emerging? What value of V is required if the beam is incident at 45°?

*If the protons are separated by a distance less than 10−14 m, they are affected by the very short range nuclear forces. Unlike gravitational and electrical forces, nuclear forces are not known precisely. However, in the study of atoms one does not need to know anything about nuclear forces beyond the fact that they are strong enough to bind together the constituent parts of the atomic nucleus.

* The forces between complete atoms are many-body forces, since the force between two atoms does depend on where other atoms are situated. The many-body nature of the force arises because the distribution of the constituent charged particles within each atom is changed by the presence of other atoms. But if the distribution of all the charge within each atom were specified, then the total force could be found by adding the forces between each pair of charges.

*The value of e(unlike ε0) does depend on the results of measurements, and there is an uncertainty associated with it. The best value of e, adopted in 1986, is e = (1.60217733 ± 0.00000049) × 10−19 coulombs. For most practical purposes this error is so small that it is of no consequence.

*There are particles called quarks carrying charges which are fractions of e. However, quarks are never observed separately, but always combine to form particles which do not have fractional charge.

*Strictly speaking Equation (1.7) only applies to assemblies of stationary charges. The exact expression for the electric field, due to moving particles, which is discussed in Chapter 13, differs appreciably only for speeds close to the speed of light. In ordinary matter Equation (1.7) is a good approximation to the atomic electric field at a particular moment.

* Very high fields are maintained in the field emission microscope, which is described in a problem at the end of this chapter.

*A small part of the charge within δV should be ascribed to the macroscopic density ρ, which is assumed to continue smoothly right up to the boundary.

*The energy of an ionic crystal is discussed in more detail in Chapter 3 of Kittel, Introduction to Solid State Physics, 6th edition, Wiley, 1986. Here we are chiefly concerned to establish the relations (1.35) and (1.36) for the potential energy of an assembly of charges, and to find the order of magnitude of the interaction of electronic charges at spacings on an atomic scale.

CHAPTER 2

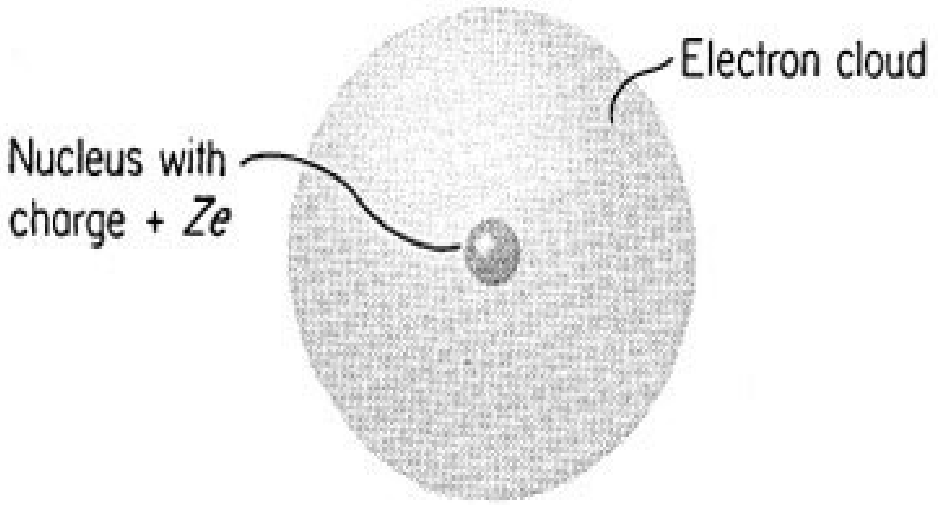
Dielectrics

The electric field generated by an assembly of many charges can be found by applying the principle of superposition. In practice it is not easy to specify the charge distribution in a piece of matter, since atomic charges are not fixed in position. Any one charge moves about under the action of internal atomic forces and the forces due to external fields. We have already seen that when a metallic conductor is placed in an external field, conduction electrons move until induced charges ensure that the macroscopic field is zero everywhere inside the metal. In an insulator, all the electrons are bound to particular atoms, but external fields still slightly displace the electrons in each atom. The displacement results in the appearance of induced charges which reduce the field in the insulator, though not completely cancelling it. The purpose of this chapter is to discuss the induced charges occurring in insulating materials, or dielectrics as they are often called, and to formulate the rules obeyed by the electrostatic field in the presence of dielectrics.

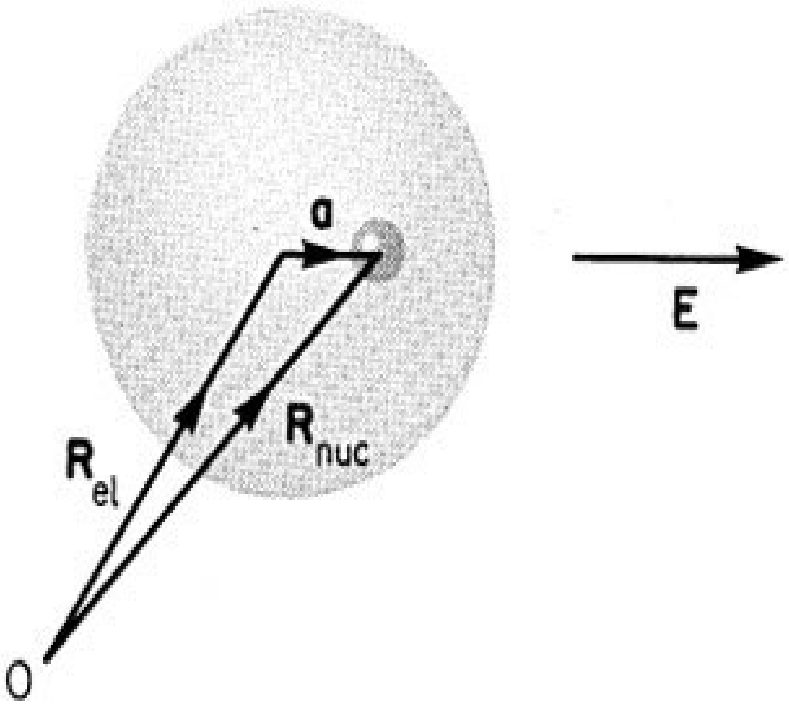
2.1 POLARIZATION

The isolated neutral atom shown in Figure 2.1(a) has Z electrons moving around a nucleus carrying a charge + Ze. (The number Z, which is called the atomic number of the atom, determines its chemical behaviour: each chemical element has a different atomic number, and in the periodic table the elements are listed in the order of increasing atomic number.) In Figure 2.1(a) the size of the nucleus is enormously exaggerated; its diameter in a real atom is only about one ten thousandth of the atomic diameter. The electrons are represented as being smeared out in a cloud around the nucleus, since according to quantum mechanical theory this is the most useful way of thinking of the electrons when we are discussing their electrostatic properties.

Figure 2.1. Polarization of an atom by an electric field.



(a) An isolated atom of the element with atomic number Z.



(b) The electric field displaces the electron cloud.

In the absence of an external field, the nucleus of an atom is at the centre of the electron cloud, where it experiences no net electrostatic force. It is in a position of stable equilibrium, so that if the nucleus were displaced from the centre of the electron cloud a restoring force would arise from the mutual attraction of nucleus and electrons. What happens if the atom is placed in an external electric field pointing towards the right as in Figure 2.1(b)? The field pushes the positively charged nucleus to the right and the negatively charged electrons to the left. As a result, the charges move until a new equilibrium is established. Most of the mass resides in the nucleus, and if the centre of mass of the atom remains at rest, the electrons move away from the nucleus. This is illustrated schematically in Figure 2.1(b). Once again the diagram is not at all to scale. The internal atomic field is much stronger than any possible external field. Consequently the distortion of the electron cloud is extremely small, and would not be visible on a scale drawing. The little movement which does occur is mostly near the outside of the atom, where the atomic field is least strong, so that the electron cloud is no longer spherically symmetrical. The net effect, however, is that the centre of the electron distribution is shifted away, and no longer coincides with the centre of the nucleus. In its new equilibrium position the atom is said to be polarized.

Figure 2.1(b) represents an electron cloud in which the centre of mass has been displaced a distance a. The vector

(2.1)

which points in the direction of the electric field, is called the dipole moment of the polarized atom. We have already met dipole moments in section 1.5.3 in connection with an electric dipole made up of two point charges ± q separated by a distance a. The magnitude of the dipole moment of the electric dipole was defined as qa, consistently with Equation (2.1). The dimensions of dipole moment are [charge] [length] and dipole moments are measured in coulomb metres.

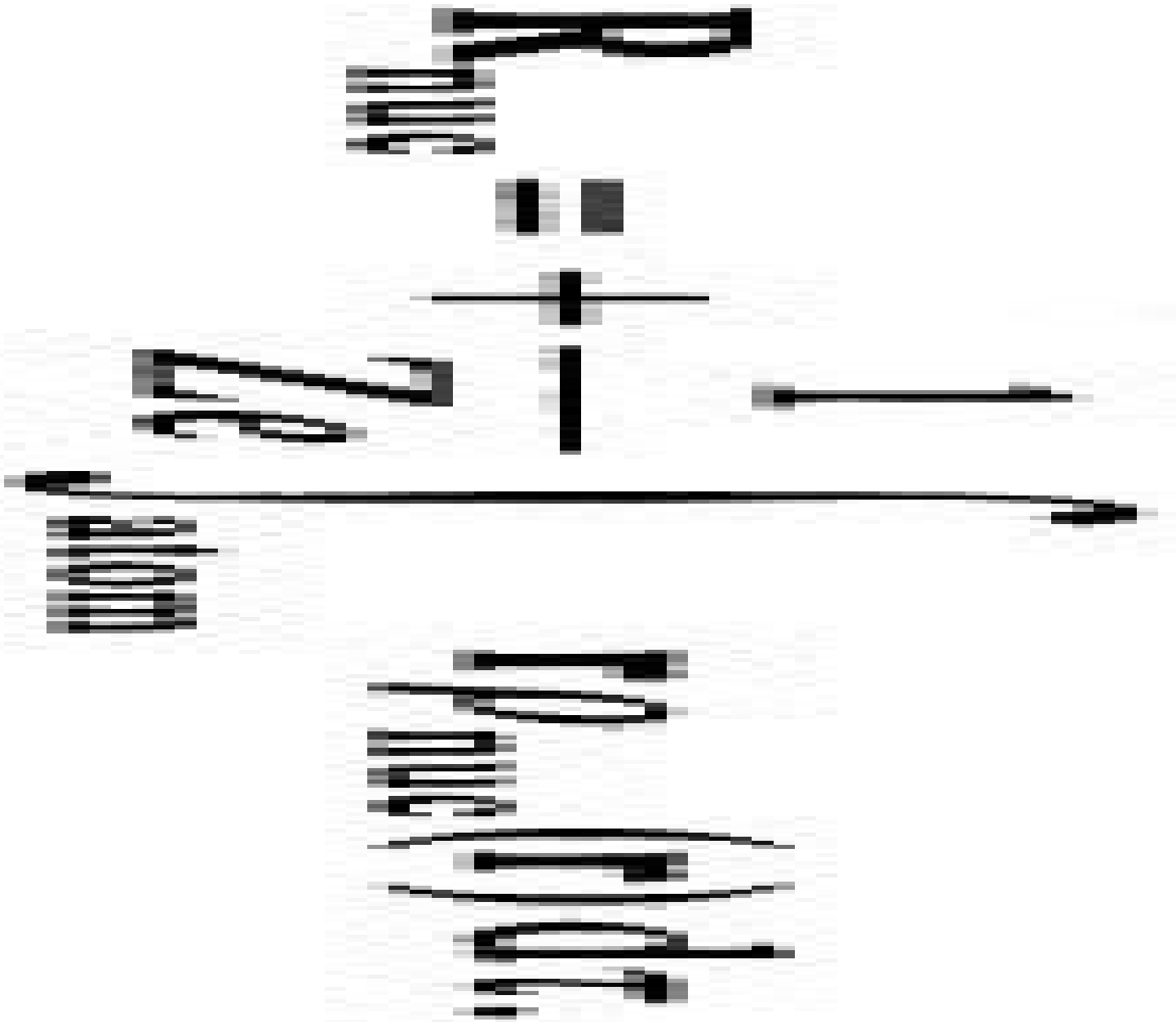
The dipole moment of a polarized atom can be expressed concisely in terms of its atomic charge density. If the charge density of the electron cloud is $\rho_{el}(r)$, referring the position vector r to the origin at O in Figure 2.1(b), then the total charge carried by the Z electrons is

The position vector Rel of the centre of mass of the electron cloud (i.e. the mean position vector, averaged over the charge distribution) is

$$R_{el} = \text{mean } \mathbf{r} = \frac{\int_{atom} \mathbf{r} \rho_{el}(\mathbf{r}) d\tau}{\int_{atom} \rho_{el}(\mathbf{r}) d\tau}$$

$$= -\frac{1}{Ze} \int_{atom} \mathbf{r} \rho_{el}(\mathbf{r}) d\tau.$$

Similarly, writing the charge density of the nucleus as ρnuc(r), the position vector Rnuc of the centre of the nucleus is



Substituting in Equation (2.1),

(2.2)

$$\mathbf{p} = Zea = Ze(\mathbf{R}_{nuc} - \mathbf{R}_{el})$$

$$= \int_{atom} \mathbf{r}(\rho_{nuc}(\mathbf{r}) + \rho_{el}(\mathbf{r}))\,d\tau$$

$$= \int_{atom} \mathbf{r}\rho_{atomic}(\mathbf{r})\,d\tau,$$

where ρatomic(r) is the total atomic charge density including the contributions of both the nucleus and the electron cloud.

When an insulator is placed in an electric field, all its atoms become polarized. For simplicity, consider a dielectric material consisting only of the elememt with atomic number Z, and containing N atoms per unit volume. In a uniform electric field, the centre of mass of the electron cloud in each atom moves the same distance away from the nucleus, in a direction opposite to the electric field. Each atom therefore acquires a dipole moment of magnitude Zea in the direction of the field*. The displacement of electrons leads to the appearance of induced charges on the surface of the dielectric. We will now relate the dipole moments of the individual atoms to the density of the surface charge—the polarization charge as it is called.

The appearance of a surface charge density means that a macroscopic volume at the surface carries a net charge. But the macroscopic charge density inside a uniformly polarized dielectric remains zero. Imagine a small cubical box inside the dielectric before it is polarized, as shown in section at the top of Figure 2.2. The box is small on the macroscopic scale, but large enough to contain many atoms (the box should really contain far more atoms than have actually been drawn). Some atoms are partly inside and partly outside the box, but on average such a box contains no net charge. When an electric field is switched on, pointing from left to right, each atom becomes polarized, and all the electron clouds move to the left. On average the electrons move a distance a, and in the schematic diagram this shift is represented by the shaded section of width a. Exactly as much charge leaves the left-hand side of the box as enters the right-hand side, and the box remains electrically neutral. Notice that this is only true for uniform polarization. If the polarization is non-uniform, different amounts of charge pass through opposite faces of the cube; we shall treat this problem in section 2.3.

Even with uniform polarization, neutrality is not preserved at the dielectric

surface. The box in the third part of Figure 2.2 encloses an area δS on the left-hand surface of the dielectric. When the dielectric is polarized, the shaded charge enters the box, but no charge leaves it. The volume of the shaded zone is $a\delta S$ and the average charge density of the electrons is $(-NZe)$ so the box acquires a net charge $(-NZea)\delta S$. In other words, the electric field induces a polarization charge of surface density $\sigma p = -NZea = -Np$. Similarly, the box on the right-hand surface loses electrons and acquires a net positive charge, corresponding to a surface charge density $(+Np)$.

We can write the charge density on both surfaces in the same form by using vector notation. The dipole moment p is a vector pointing in the same direction as the field E. Referring again to Figure 2.2, we see that if δS is an outward normal to the dielectric surface, in both cases the polarization charge on the area δS can be written as

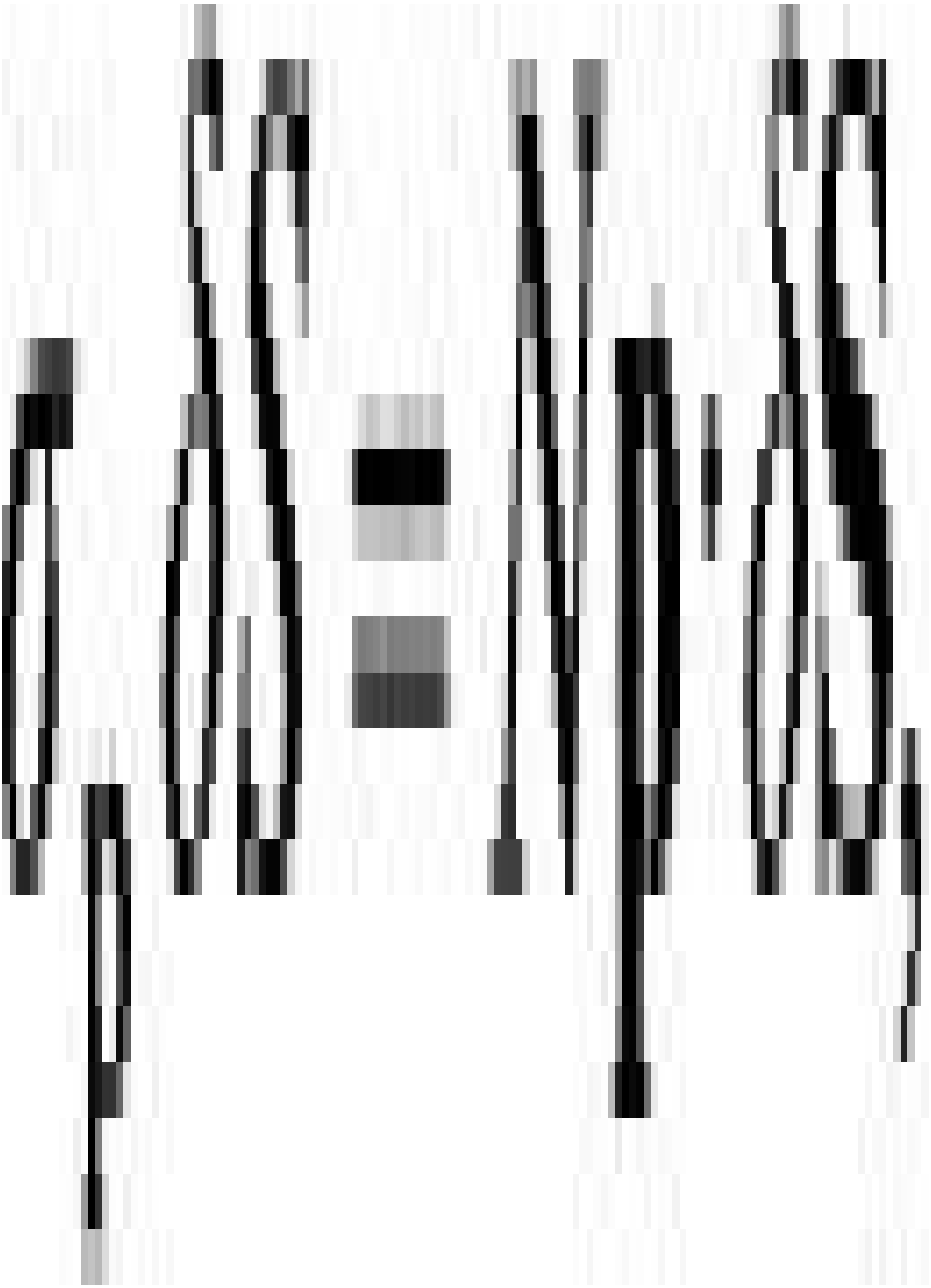
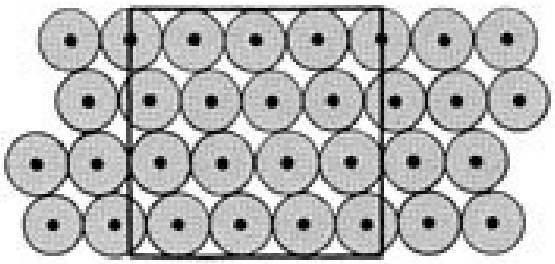
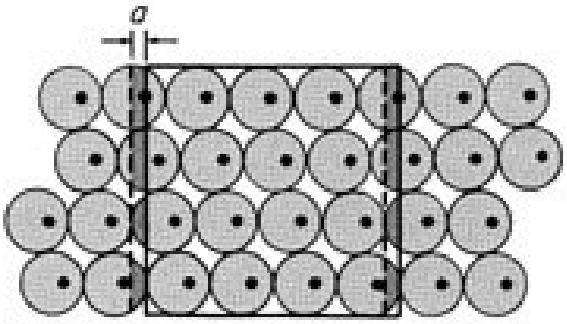


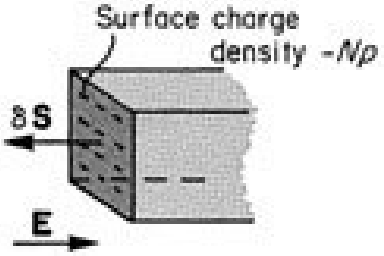
Figure 2.2. The surfaces of a polarized dielectric carry a charge, even though all the atoms in the dielectric remain electrically neutral.

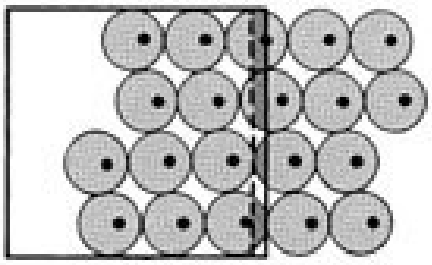


No field

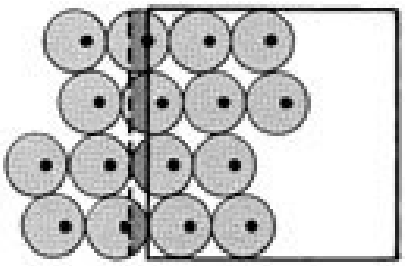


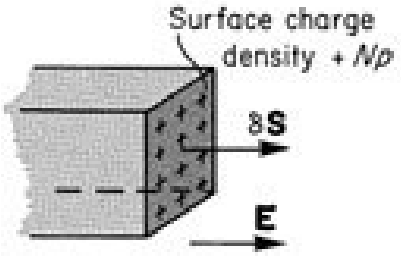
A small volume embedded in the dielectric remains neutral when the atoms are polarized by the uniform field **E**.





If the field points from left to right, a volume enclosing part of the left-hand surface *gains* electrons and acquires a net negative charge .....





while a volume enclosing part of the right-hand surface *loses* electrons and acquires a net positive charge.

or

(2.3)

where

(2.4)

The vector P, which represents the dipole moment per unit volume of the dielectric, is called the polarization.

Equation (2.3) holds generally, even for surfaces which are not normal to the electric field. In Figure 2.3 the surface element δS is at an angle θ to the field, and its projected area along the field is $\delta S \cos \theta$. The net charge acquired by the box after polarization is

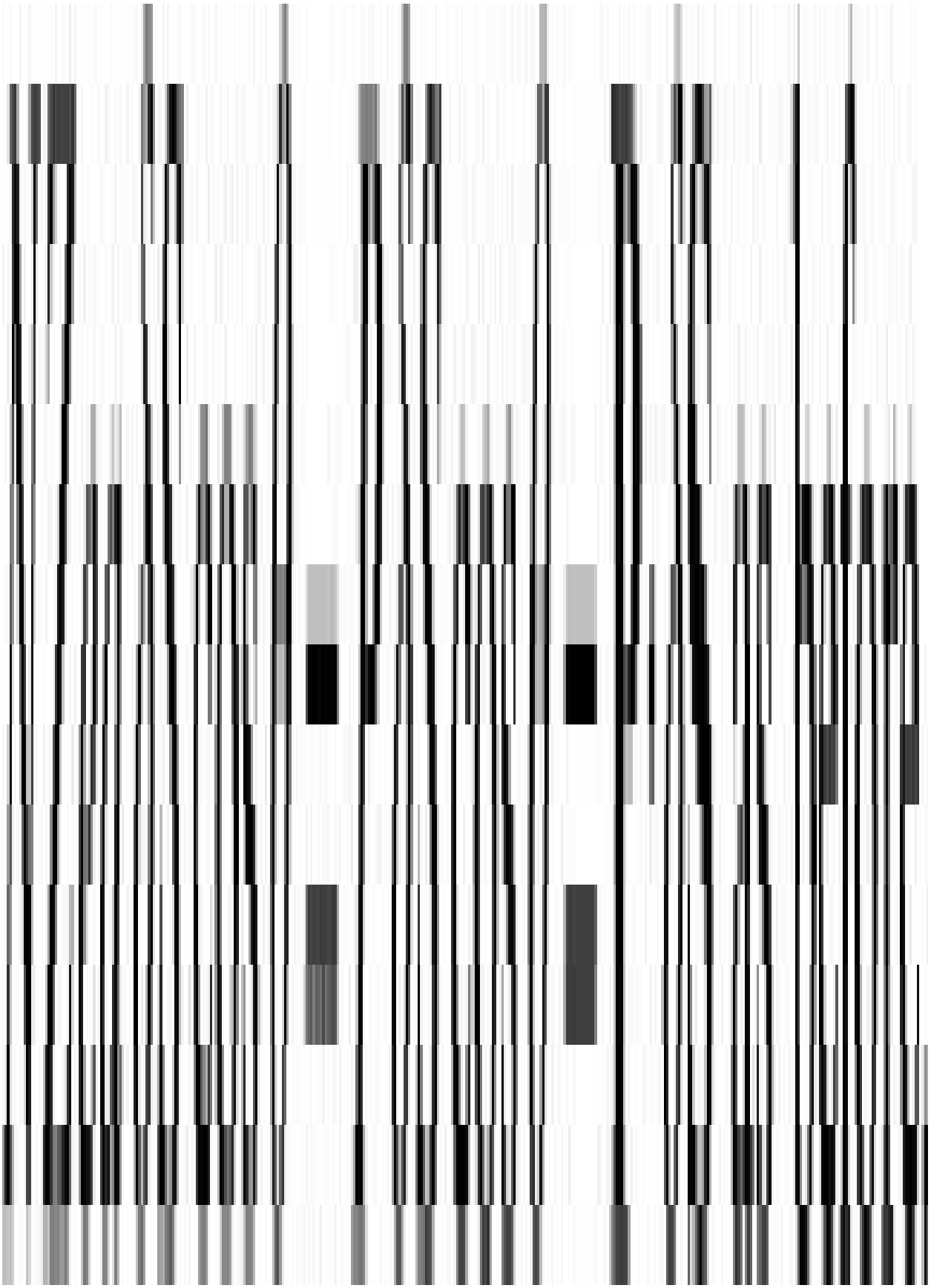
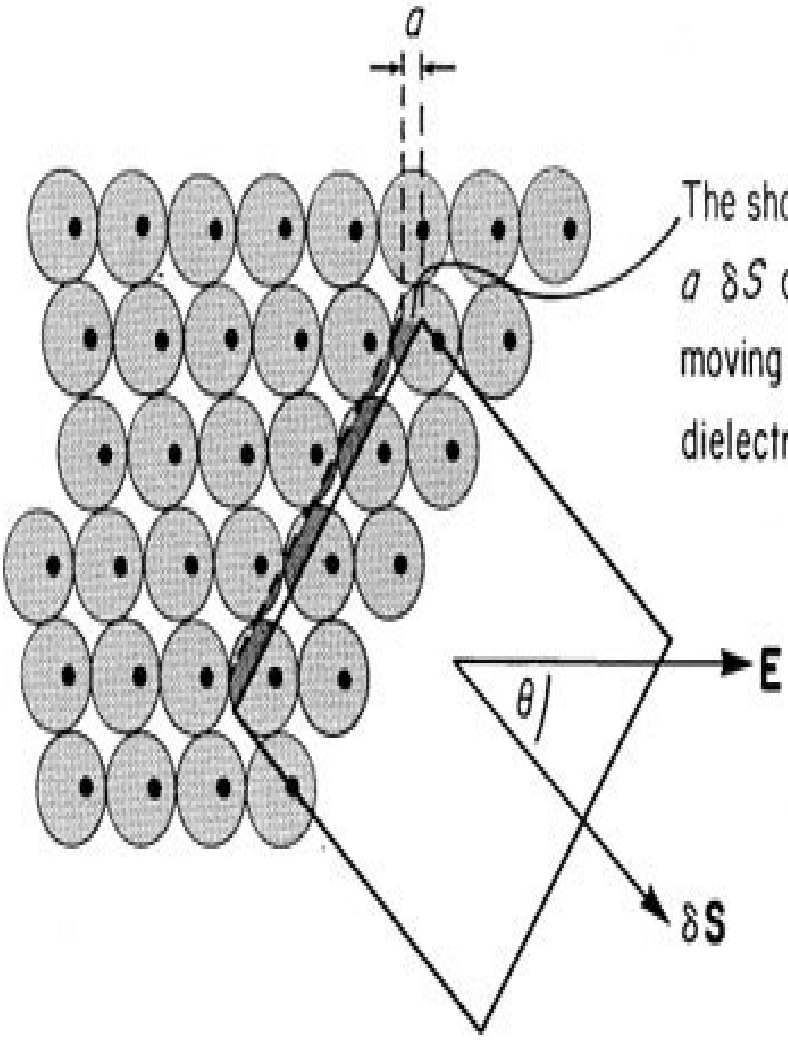
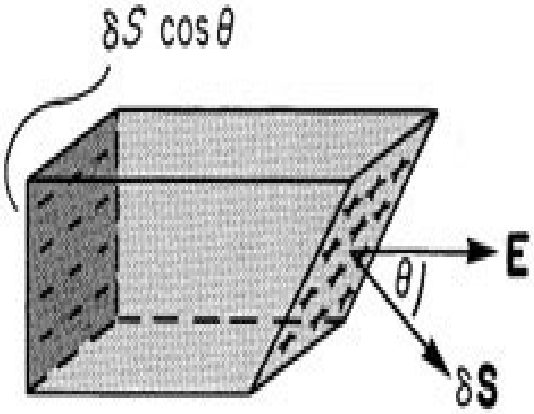


Figure 2.3. The polarization charge density is reduced on a surface inclined to the electric field.



The shaded electrons within a volume $a\ \delta S \cos\theta$ represent the net charge moving out of the box when the dielectric is polarized.



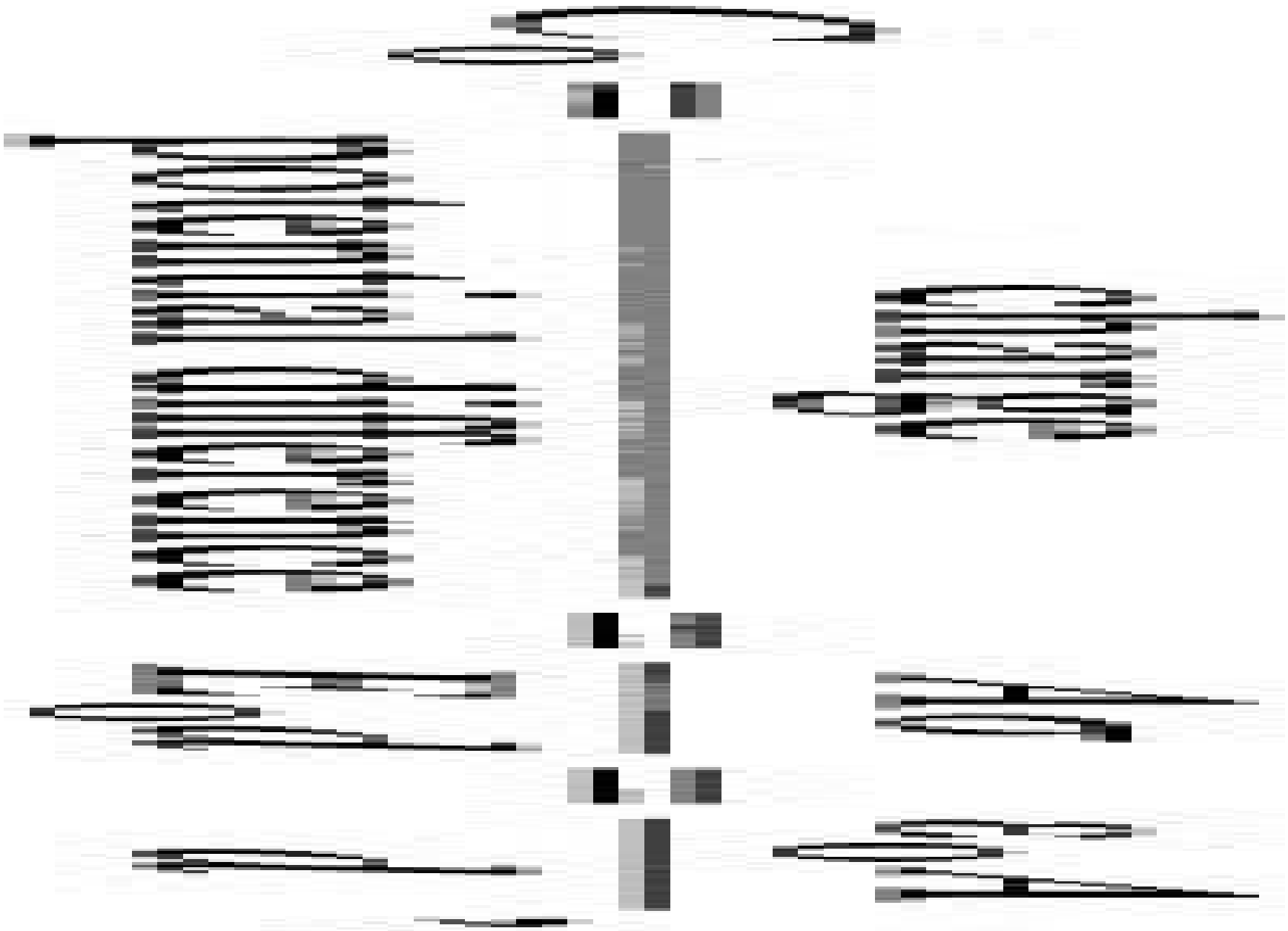
In this piece of electrically-neutral material, the negative charge on the face perpendicular to the electric field is balanced by an equal and opposite charge spread over the larger area of the inclined face.

Although for simplicity we have been considering a solid made up of identical atoms, the whole argument applies equally well to any material. Equation (2.3) still applies, with P representing the total dipole moment per unit volume including contributions from all the different kinds of atom which are present.

2.2 RELATIVE PERMITTIVITY AND ELECTRIC SUSCEPTIBILITY

Polarization charges induced on the surface of a dielectric material make a contribution to the macroscopic electric field inside the material. Their contribution is called the ‘depolarizing field’, because the sign of the induced charge always ensures that the field just inside the dielectric surface is less than the field just outside. This is illustrated in Figure 2.4, which shows a slab of dielectric material in a uniform field. The depolarizing field points away from the positive polarization charges towards the negative charges, i.e. from right to left, thus partly cancelling the external field. The net macroscopic field is therefore less inside the dielectric than outside: this is represented in the figure by the lower density of field lines inside the dielectric.

As a result of the presence of the depolarizing field, the capacitance of a capacitor is increased by filling it with dielectric material. This can be understood by considering a parallel plate capacitor like the one shown in section in Figure 2.5. Suppose that the plates each have an area A and that they are separated by a distance d. When the capacitor is in vacuum, and there are surface charges $\pm \sigma$ A on the plates, we can find the field E0 between them by applying Gauss’ law to the surface S enclosing the charge $+ \sigma A$ on the left-hand plate. The flux out of S is σA, and from Gauss’ law $E0 = \sigma/\varepsilon 0$, a result already proved in section 1.4.2. The capacitance in vacuum is



the result already given in Equation (1.41).

Figure 2.4. Electric field lines in a slab of polarized dielectric. Lines end on negative polarization charges and start on positive ones. The density of lines is reduced inside the dielectric, corresponding to a smaller field inside than outside.

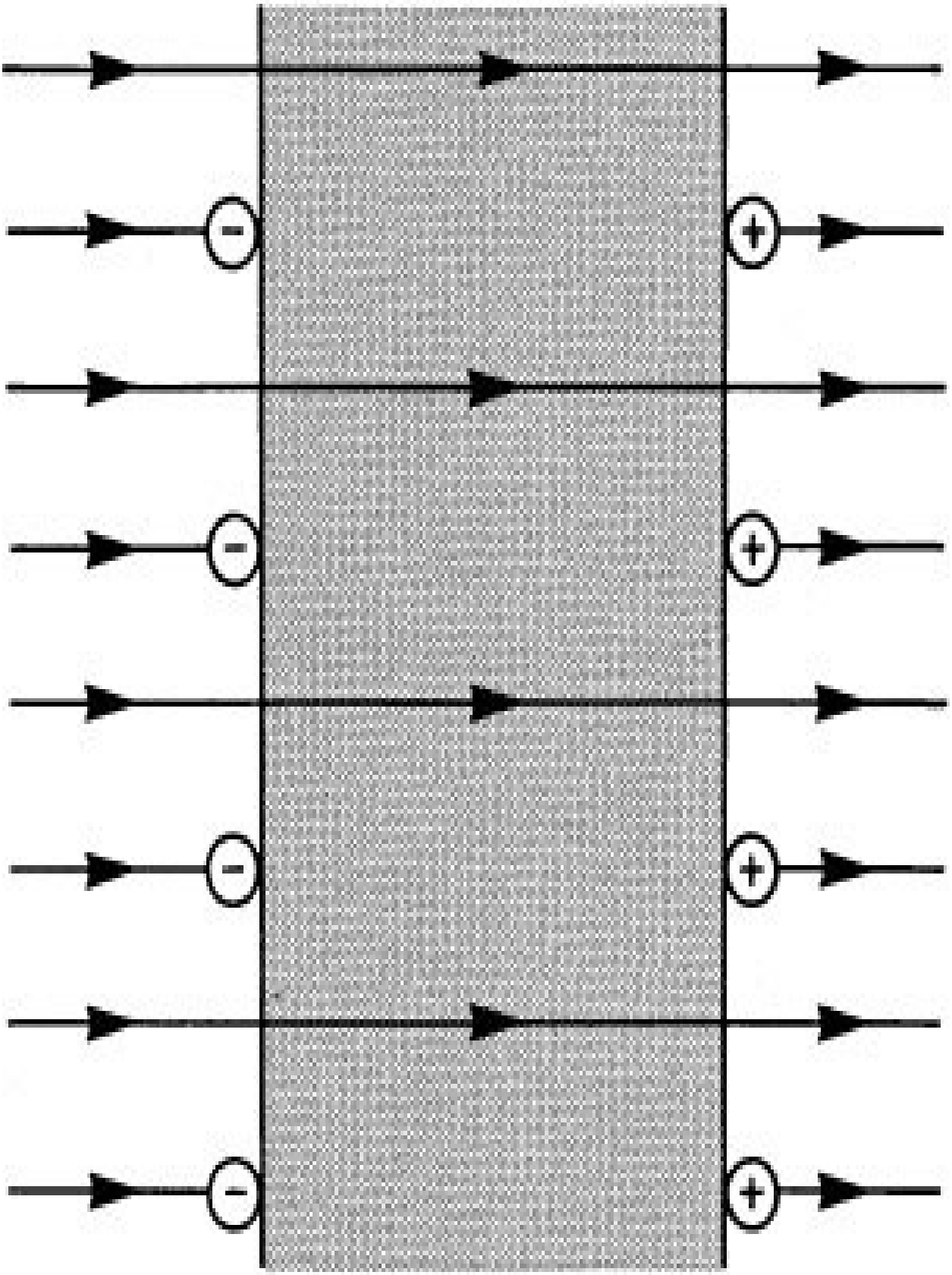
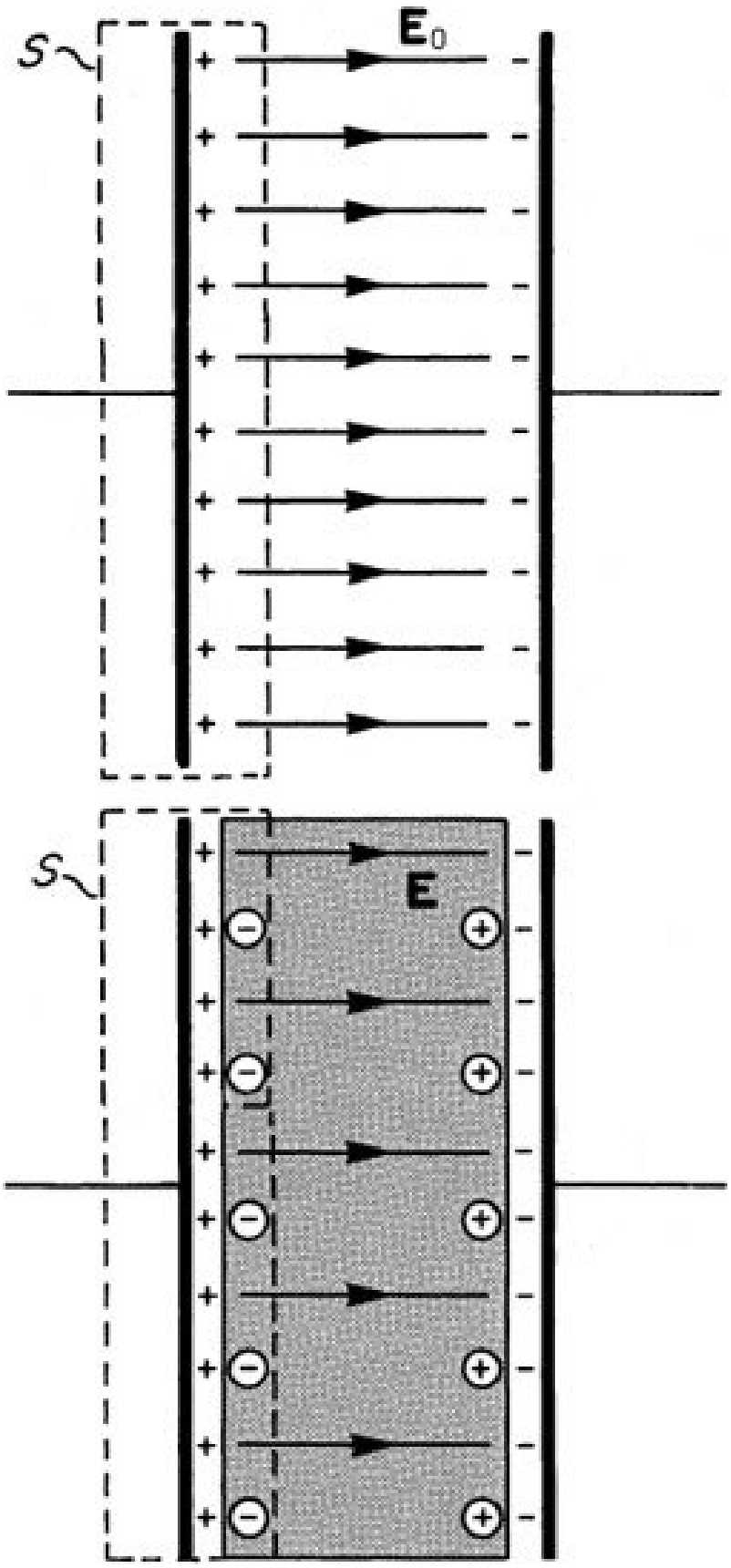


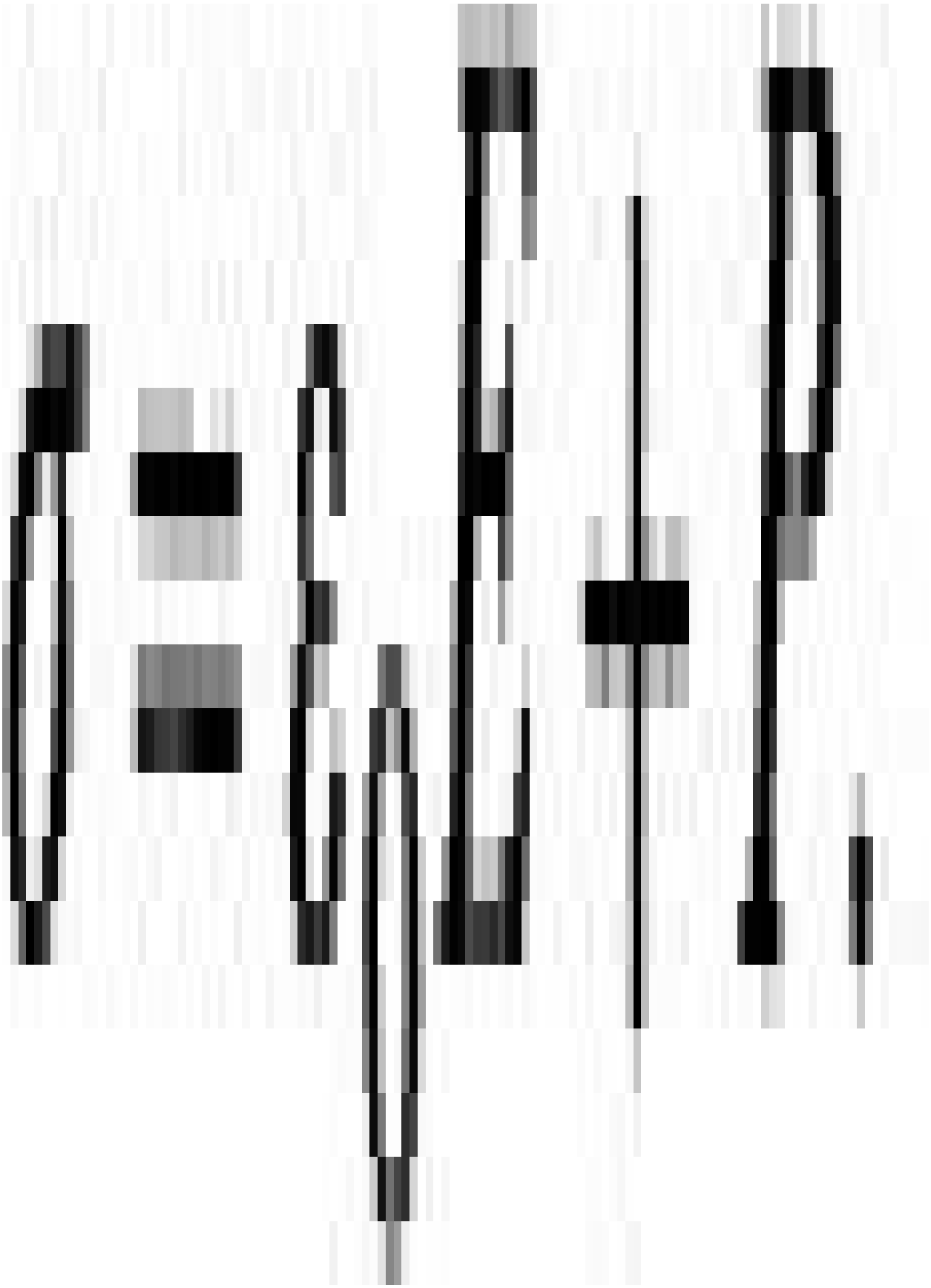
Figure 2.5. When a parallel plate capacitor is embedded in a dielectric material, the electric field between the plates is reduced. The diagram shows the field to be halved in the presence of the dielectric, corresponding to $\varepsilon = 2$. To show the charges clearly, a gap has been drawn between the dielectric and the plates.


S
E_0
S
E

Now keeping the same charge densities ± σ on the conducting plates, fill the space between them with dielectric material. The dielectric acquires a polarization P, and induced charges appear on its surface. As shown in the lower part of Figure 2.5, the area A of the dielectric which is enclosed within S carries a negative polarization charge. The magnitude of this charge is (–PA).

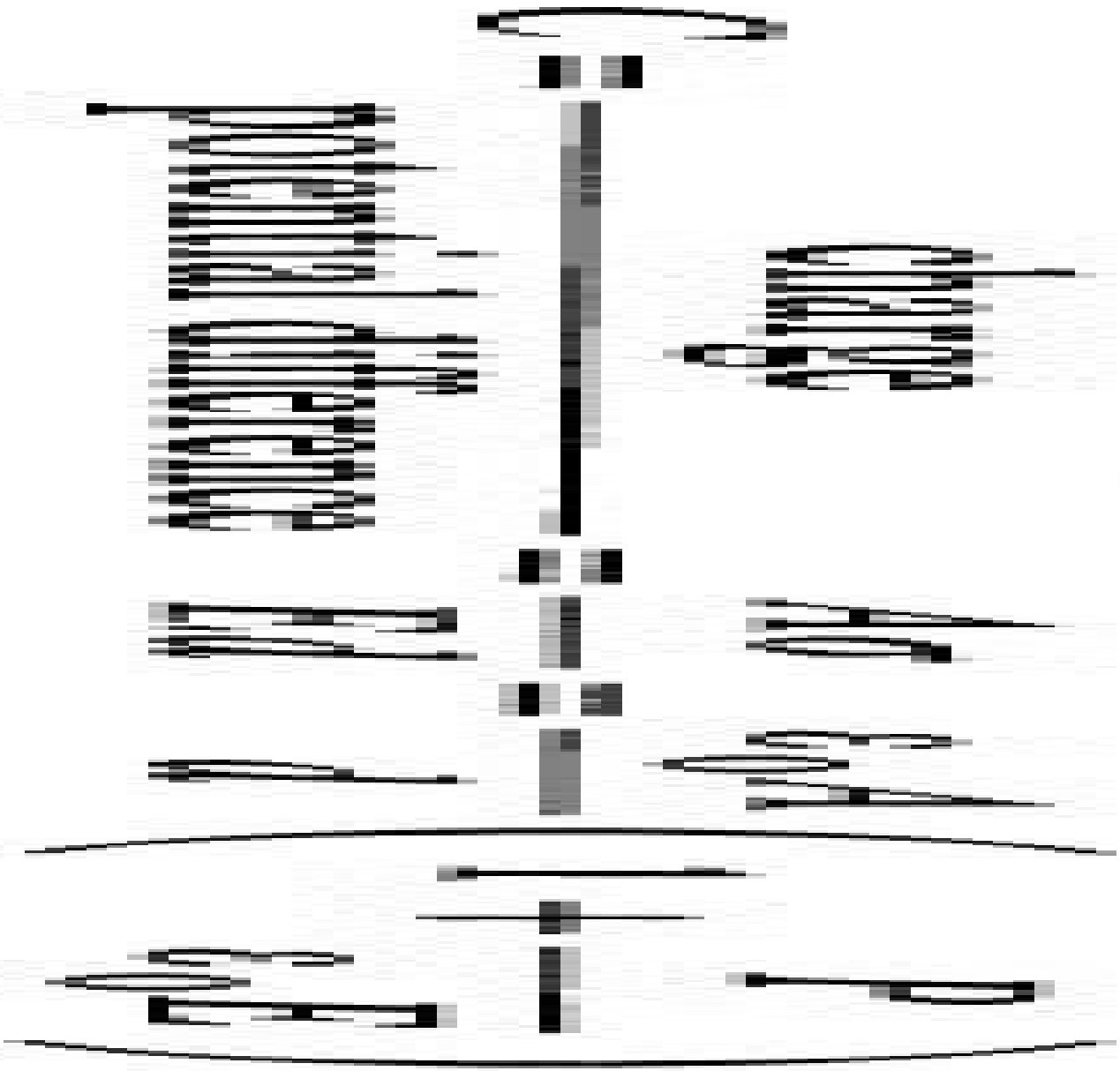
Applying Gauss' law to S, the field E between the plates satisfies

and hence

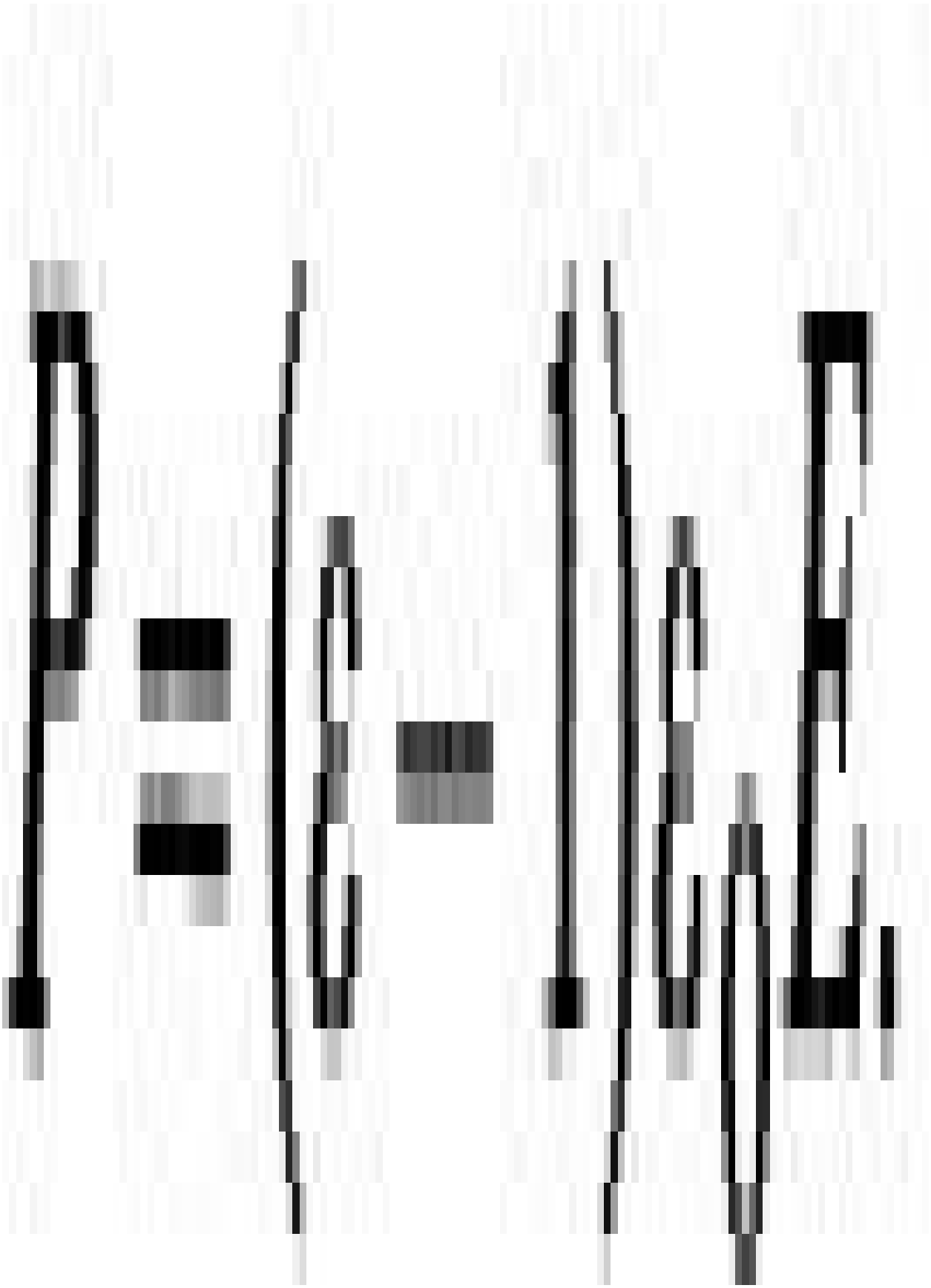


It is found experimentally that for fields below the field at which breakdown occurs, the charge and voltage on the capacitor are still proportional to one another. Furthermore, the factor ε by which the capacitance increases depends only on the nature of the dielectric and not on the size and shape of the capacitor. The factor ε, which is dimensionless, is called the relative permittivity of the dielectric material*. For the parallel plate capacitor filled with dielectric the capacitance is

The potential difference between the plates is Ed, and we can also write



Comparing the two expressions for C, we find



For the isotropic dielectrics which we are considering (see the footnote on p. 52), the vectors P and E point in the same direction, and the equation can be written in vector form as

(2.5)

Since ε is a constant, it follows that the polarization P is proportional to the electric field E, and we can write

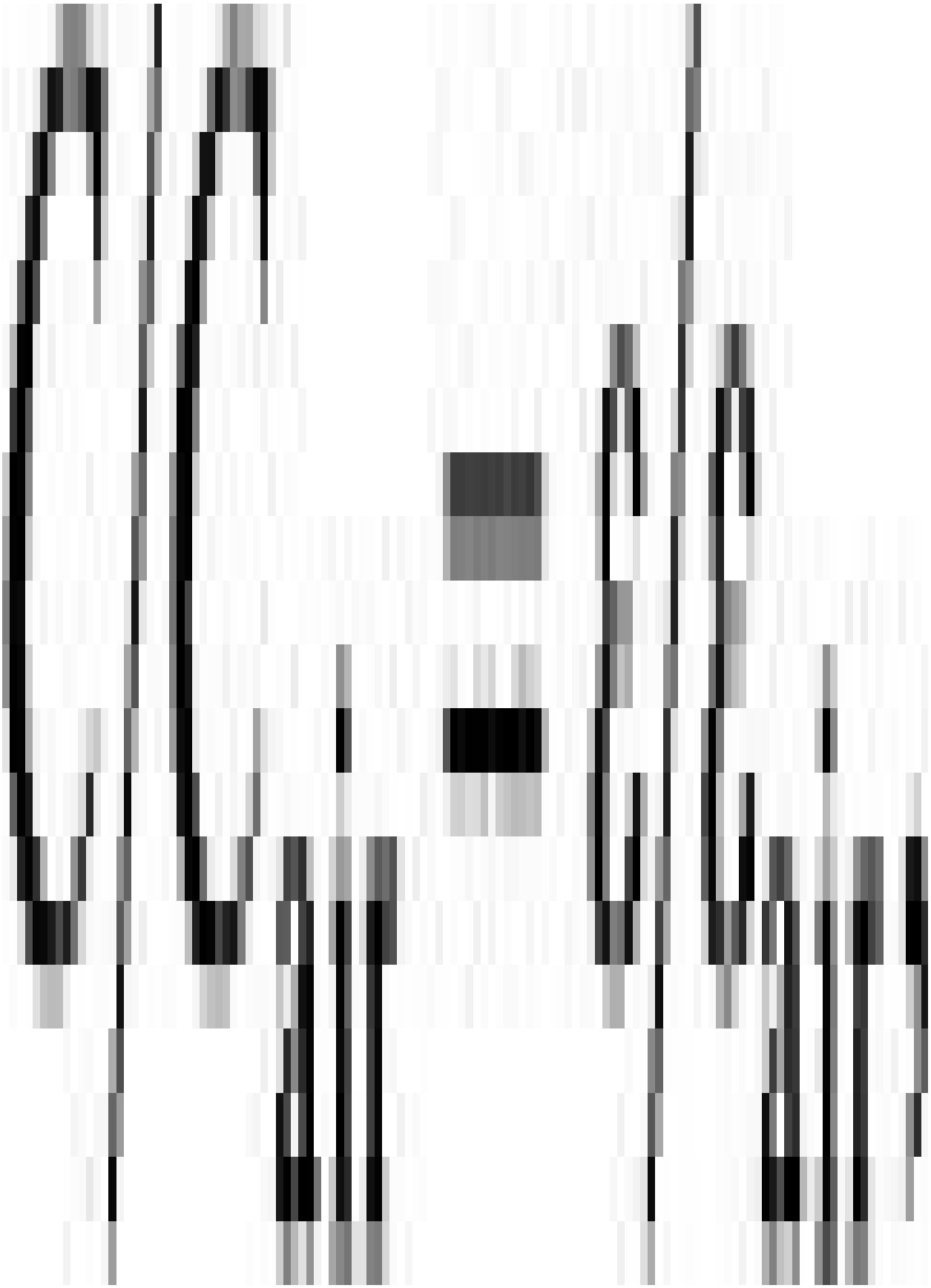
(2.6)

where the dimensionless constant of proportionality

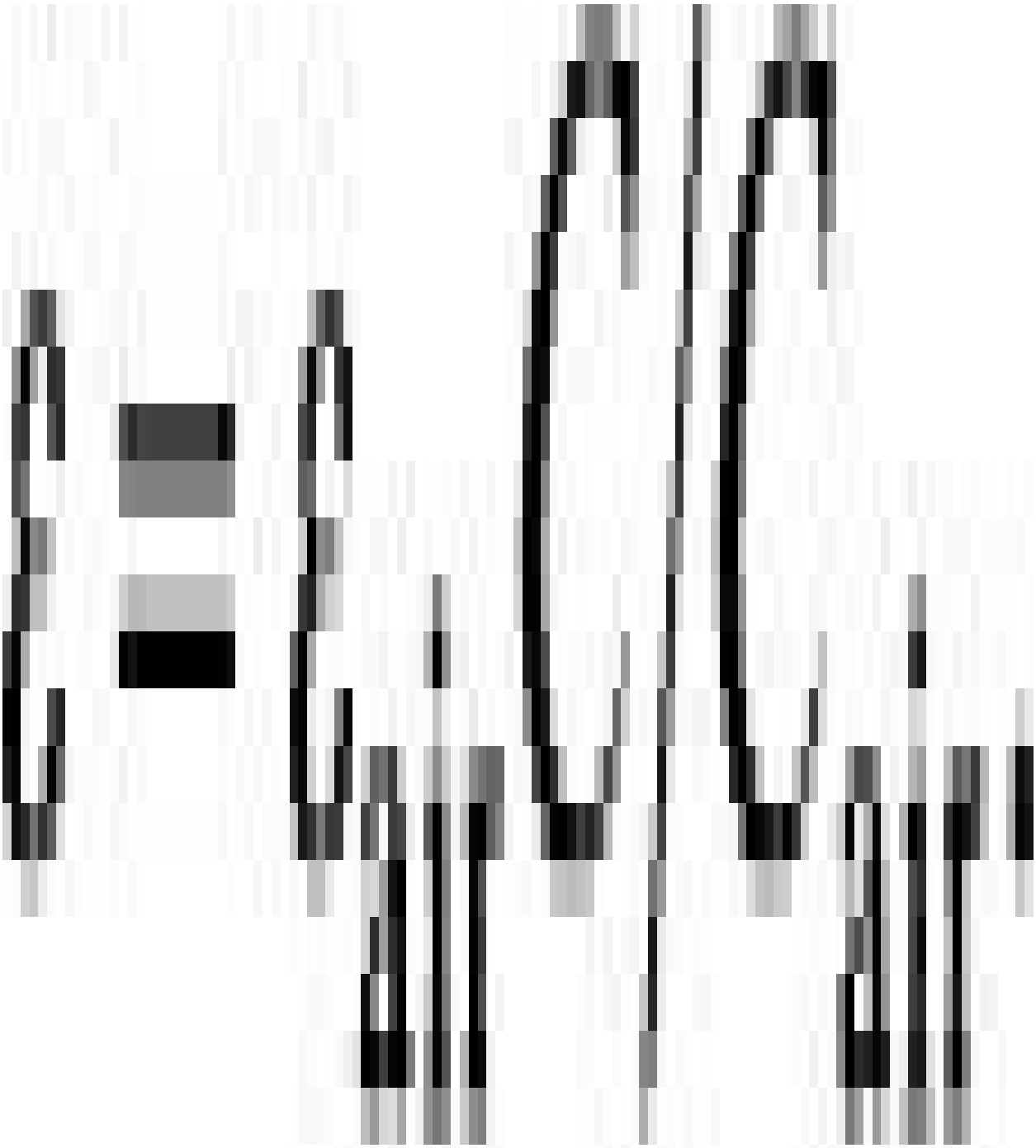
(2.7)

is called the electric susceptibility of the dielectric material. Equations (2.5) and (2.7) constitute the definitions of relative permittivity and electric susceptibility.

The relative permittivity ε of a dielectric material is in fact usually found by measuring the capacitance Cair of a capacitor in air and the capacitance C of the same capacitor when it is filled with the dielectric. The ratio of capacitances



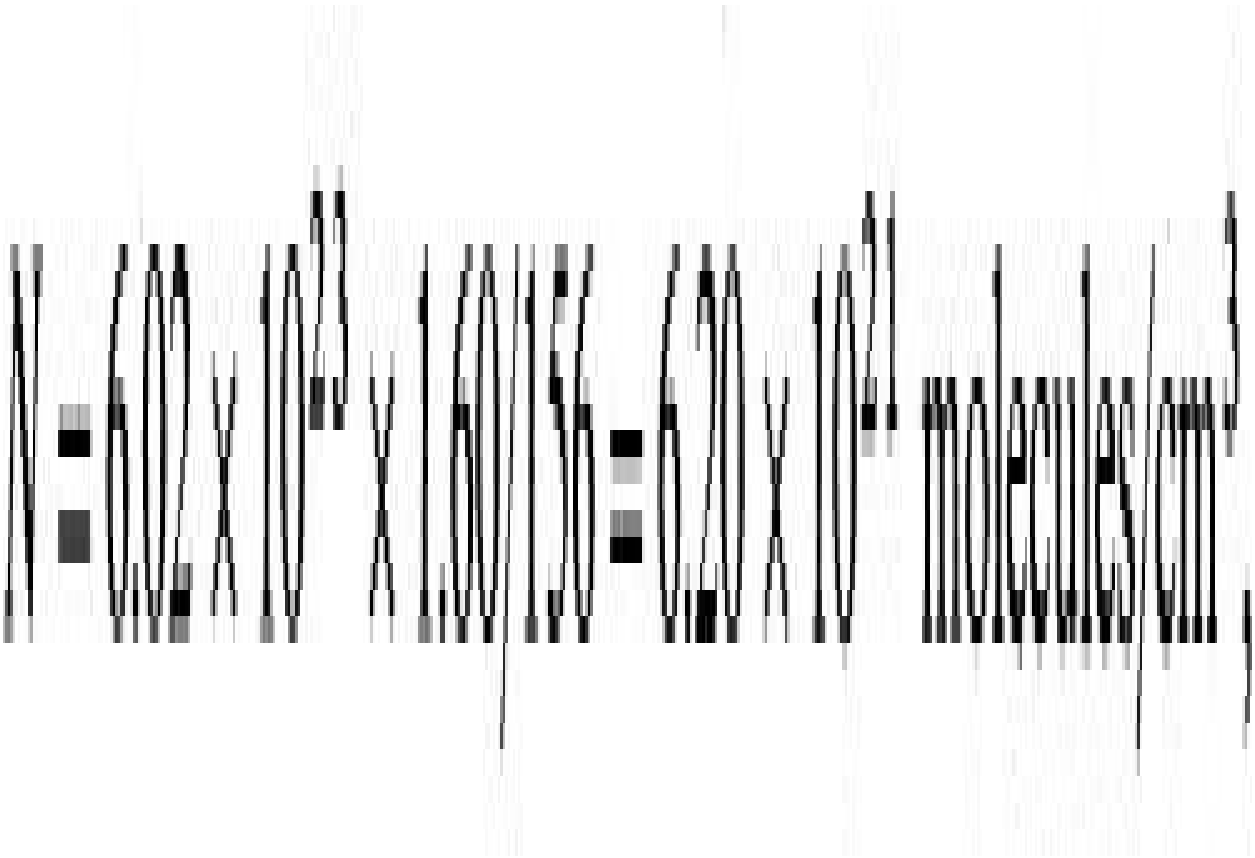
and hence



Values of relative permittivity for a number of insulators are given in Table 2.1. One sees from the table that the capacitance of a capacitor is several times larger when it is filled with a solid or liquid dielectric than in vacuum. Many of the capacitors used as components in electronic circuits are in fact made with solid dielectrics such as polystyrene in the form of thin sheets coated with conducting films.

The use of solid dielectrics gives a bonus to the value of the capacitance because the relative permittivity is larger than one, and also enables much smaller spacings to be maintained between the conductors than is possible with air-filled capacitors. By rolling up thin sheets large capacitances can be squeezed into a small volume. Typically a 1 μF capacitor capable of sustaining 50 volts across its conductors is about 1 cm in diameter and a few cm long. In contrast the capacitance of air-filled capacitors is not usually more than a few hundred pF.

Equation (2.6) always holds for isotropic dielectrics below their breakdown fields and not just in parallel plate capacitors. We can use this equation to work out the dipole moment p of an individual molecule, since the polarization $P = Np$. Let's do this for carbon tetrachloride, which at 20 °C has relative permittivity 2.24 and density 1.60gm/cm^3. The molecular weight of CC14 is 156, so that 156 gm contains 6.02×10^{23} molecules, and the molecular density is



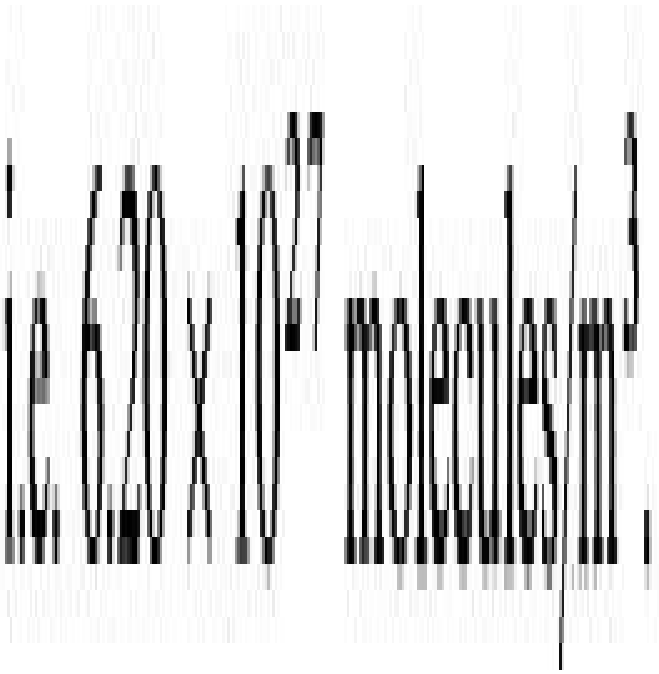
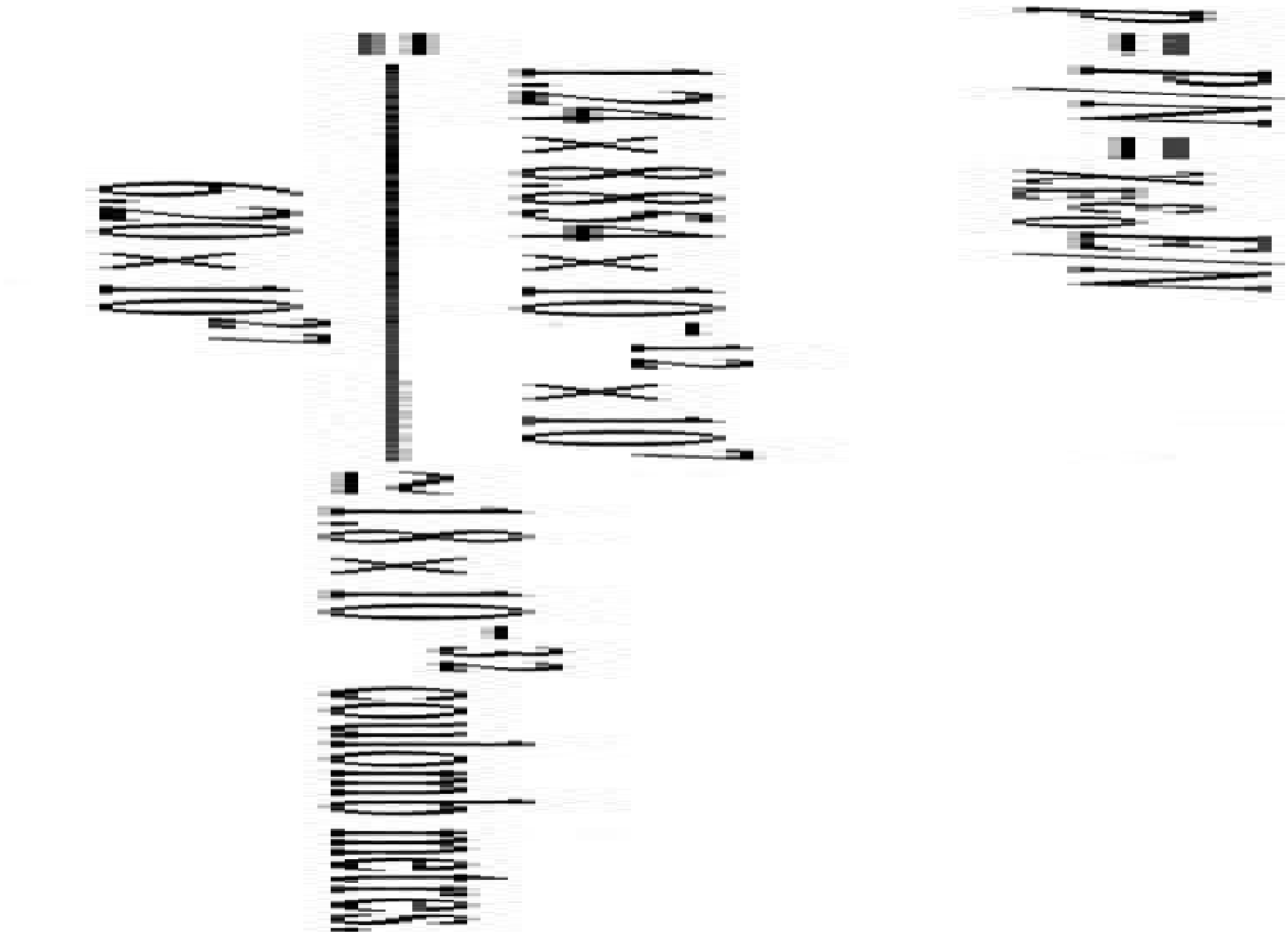


Table 2.1 The relative permittivities of insulating materials at room temperatures.

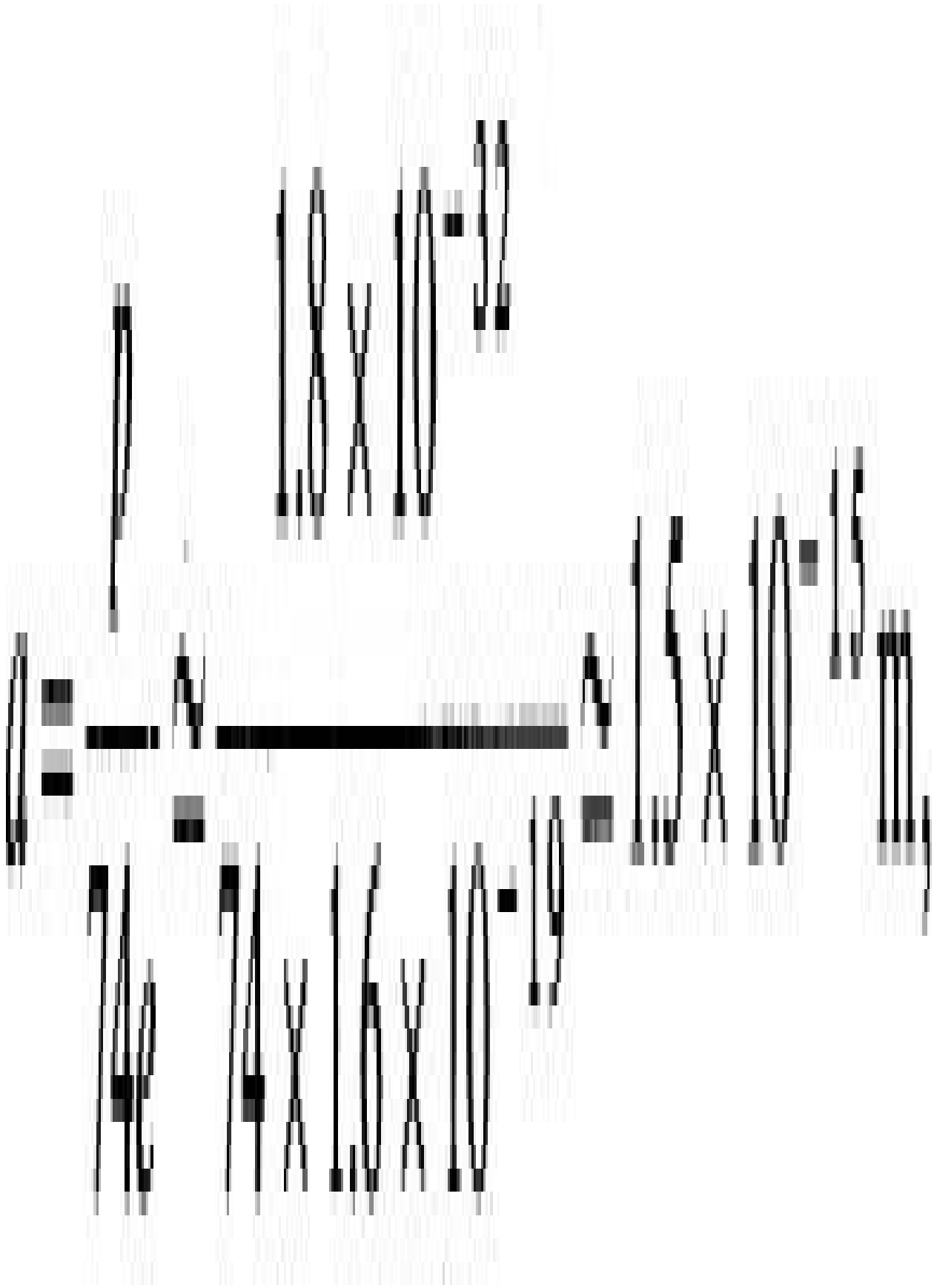
SUBSTANCE	RELATIVE PERMITTIVITY
Air at atmospheric pressure	1.00059
CCl4	2.24
Transformer oil	2.2
Paraffin wax	2–2.5
Polyethylene	2.3
Nylon	3.5
Porcelain	6
Mica	7

In a field of 10^7 volts/metre (near the highest value obtainable in a liquid before breakdown occurs), the dipole moment of single molecule is

(2.8)



There are 74 electrons in each CC14 molecule, and if the average electron displacement is a, the dipole moment of the whole molecule is p = 74 ae. In a field of 10^7 volts/metre,

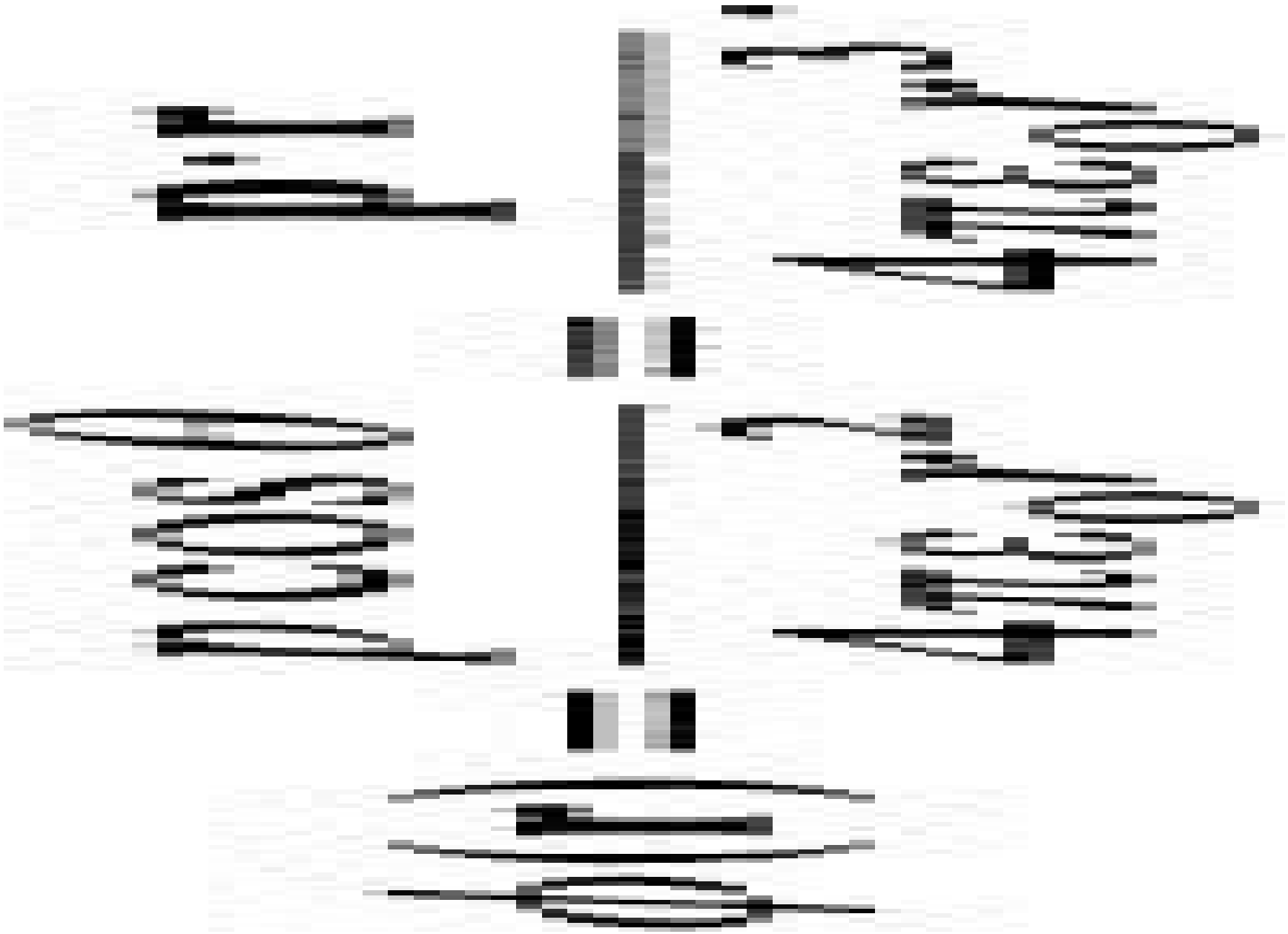


only about one hundred thousandth of the diameter of an atom. It is because of the smallness of the displacement that polarization effects are proportional to the magnitude of the electric field. The amount of polarization is determined by the balance between the external forces trying to displace electrons and the internal restoring forces. If electrons were moved away from their equilibrium position through distances comparable with the size of an atom, the restoring forces would certainly include higher powers of the displacement than the first. But the higher order terms are quite negligible for the small displacements caused by polarization, even in the highest electrostatic fields. The induced dipole moment of a molecule is always proportional to the magnitude of the electric field in which the molecule is situated.

2.2.1 The local field

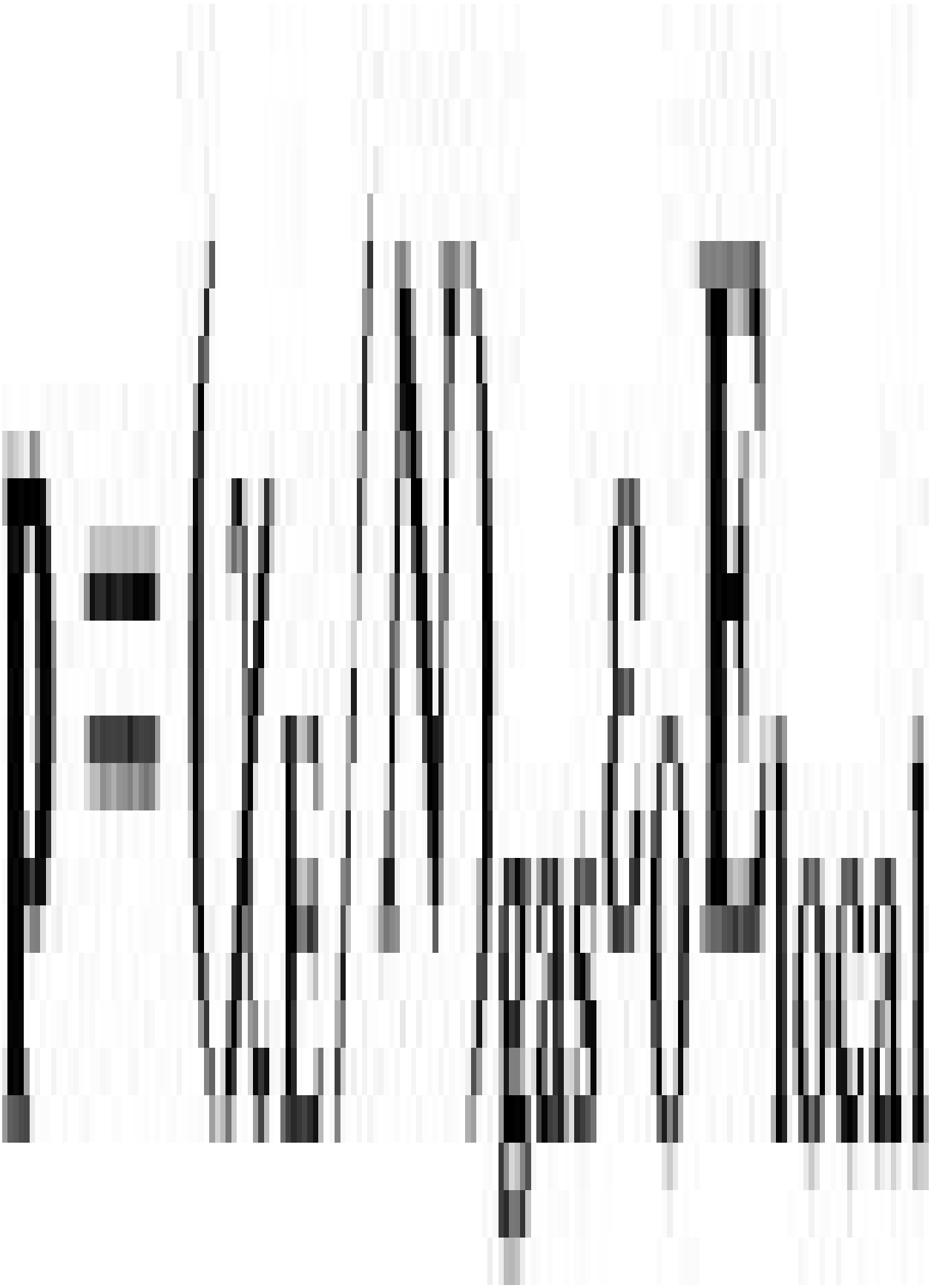
For a uniformly polarized dielectric, Equation (2.8) establishes a connection between the average dipole moment of a single molecule and the macroscopic electric field. However, the macroscopic field, which has been smoothed out over a region large enough to contain many molecules, is not necessarily the same as the local field polarizing one molecule. The local field Elocal acting on a particular molecule is the field generated by all charges outside the molecule: in other words Elocal is the same as the atomic field except that the field generated by the charges making up the molecule in question has been subtracted away.

As well as the fields due to charges outside the dielectric and to polarization charges on its surface, the local field includes contributions from nearby neutral molecules. In section 1.5.3 the potential at a distance r from an electric dipole made up of two point charges was expanded as a series in (1/r). The leading term depends on the dipole moment p, and is given in Equation (1.31) as



Exactly the same expression applies for a molecule with dipole moment p. Since this potential is proportional to $(1/r)^2$, the field associated with the dipole moment is proportional to $(1/r)^3$, and falls off rapidly with distance. The field outside a molecule also has terms in higher powers of (1/r), with magnitudes which depend on the shape of the electron clouds in the molecule. The higher order fields have an even shorter range than the dipole field, but they are sometimes important for closely neighbouring molecules.

At the site of one particular molecule in a liquid or solid dielectric, the vector sum of the short-range fields of the neighbouring molecules is usually different from zero. The local field Elocal is then equal to the macroscopic field (i.e. the slowly-varying field due to charges more than a few atomic diameters away) plus the contribution from nearby molecules. In a gas, on the other hand, molecules spend most of their time so far apart that the short-range fields are negligible, and the local field Elocal is almost the same as the macroscopic field E. When applied to a gas, Equation (2.8) thus relates the induced dipole moment p of a single molecule to the actual field Elocal = E acting on the molecule:



or

(2.9)

where

(2.10)

The constant α, called the molecular polarizability, is a property of the individual molecule: it measures the resistance of the molecule to displacement of its electron cloud. Equation (2.9) therefore holds with the same value of α even in liquids and solids, for which short-range fields are important, and in which Elocal is not the same as the macroscopic field. We shall discuss the problem of estimating the local field in liquids in section 2.2.3, but first must digress to explain the dielectric behaviour of molecules which have a permanent dipole moment, even when they are not placed in an external field.

2.2.2. Polar molecules

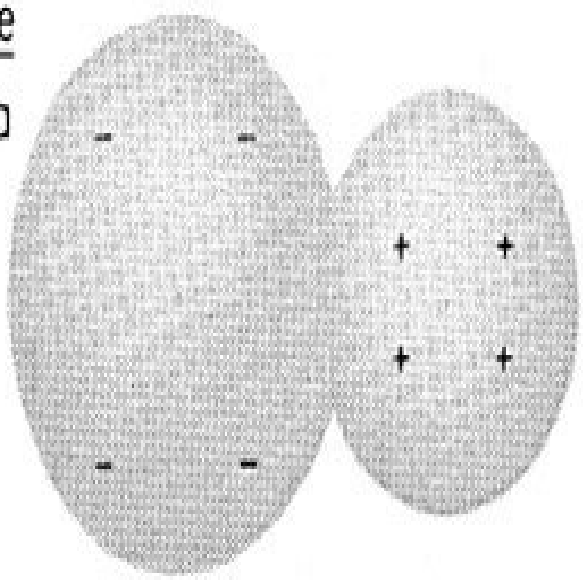
All chemical bonds are formed by the sharing of electrons between atoms. Often the sharing is unequal, and some atoms gain electron density at the expense of their neighbours. The alkali halides in the gaseous state, for example, form diatomic molecules which are almost ionic in character—just as the sodium chloride crystal discussed in section 1.5.4 contains a lattice of ions each carrying charge ±e, so the diatomic NaCl molecule consists of an Na^+ ion bound to a Cl^- ion by electrostatic attraction. Such a molecule has a permanent dipole moment

of much larger magnitude than the induced dipole moments we have been considering, since the induced dipole moments are caused by the displacement of electrons through a very small fraction of the interatomic spacing.

Figure 2.6. Sodium chloride and water molecules both have a permanent dipole moment.

Sodium chloride

Cl atom with a net negative charge

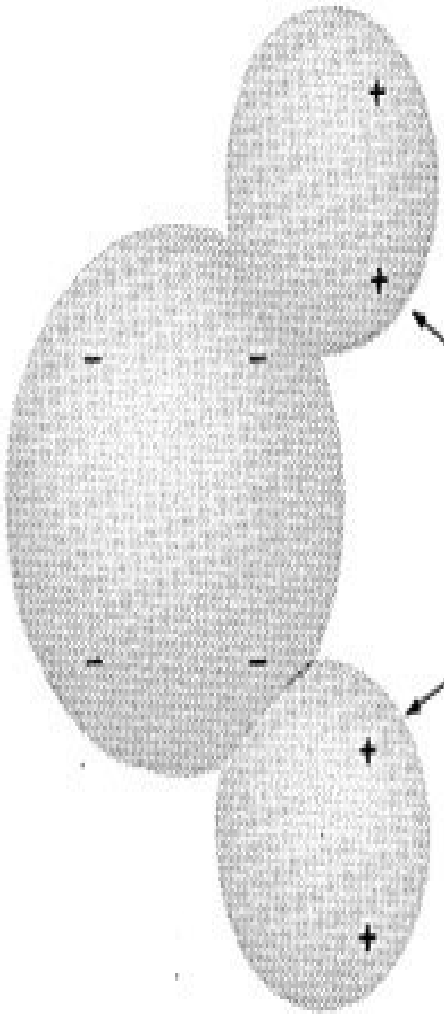


Na atom with a net positive charge

Direction of the dipole moments

Water

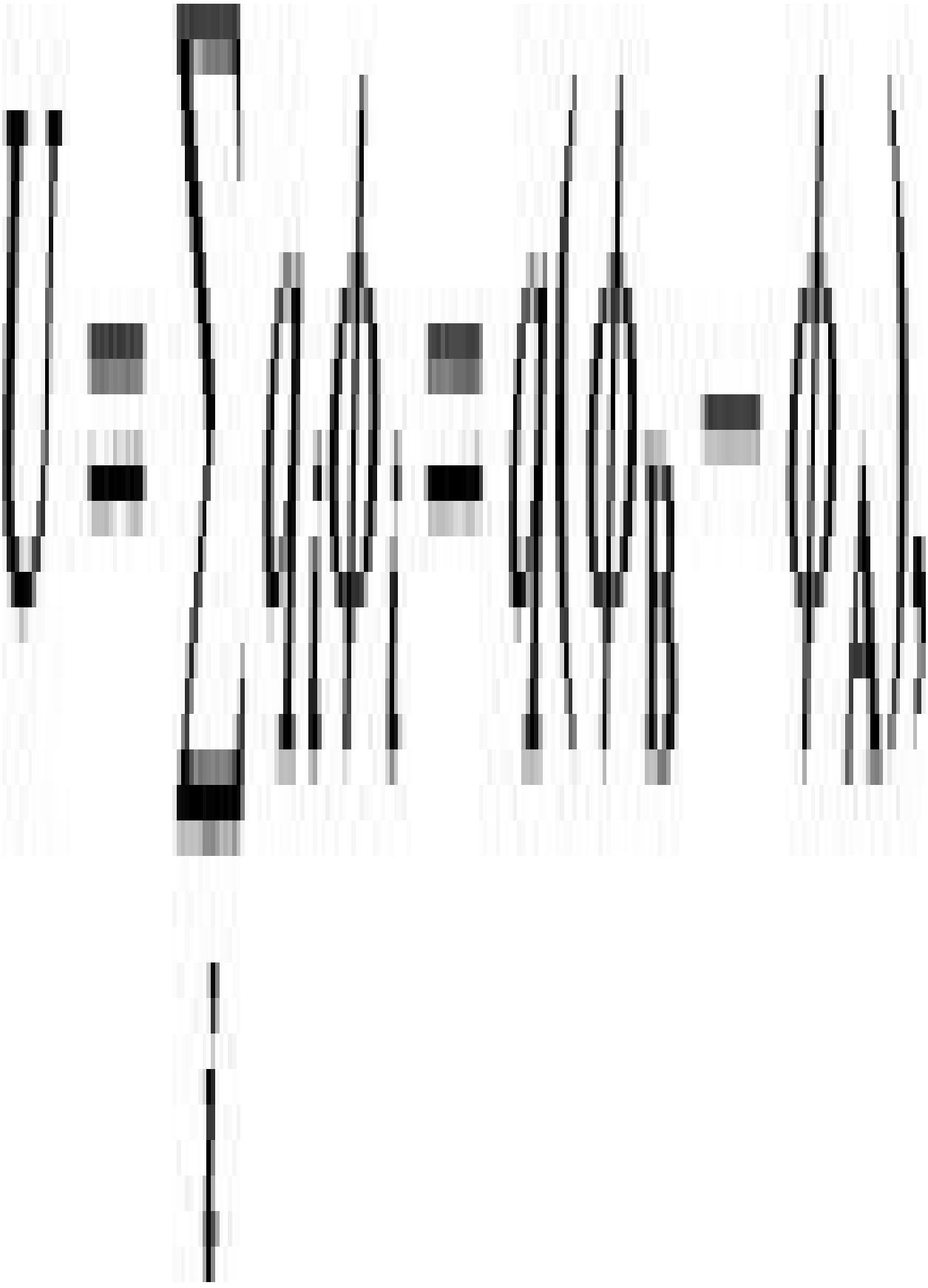
O atom with a net negative charge



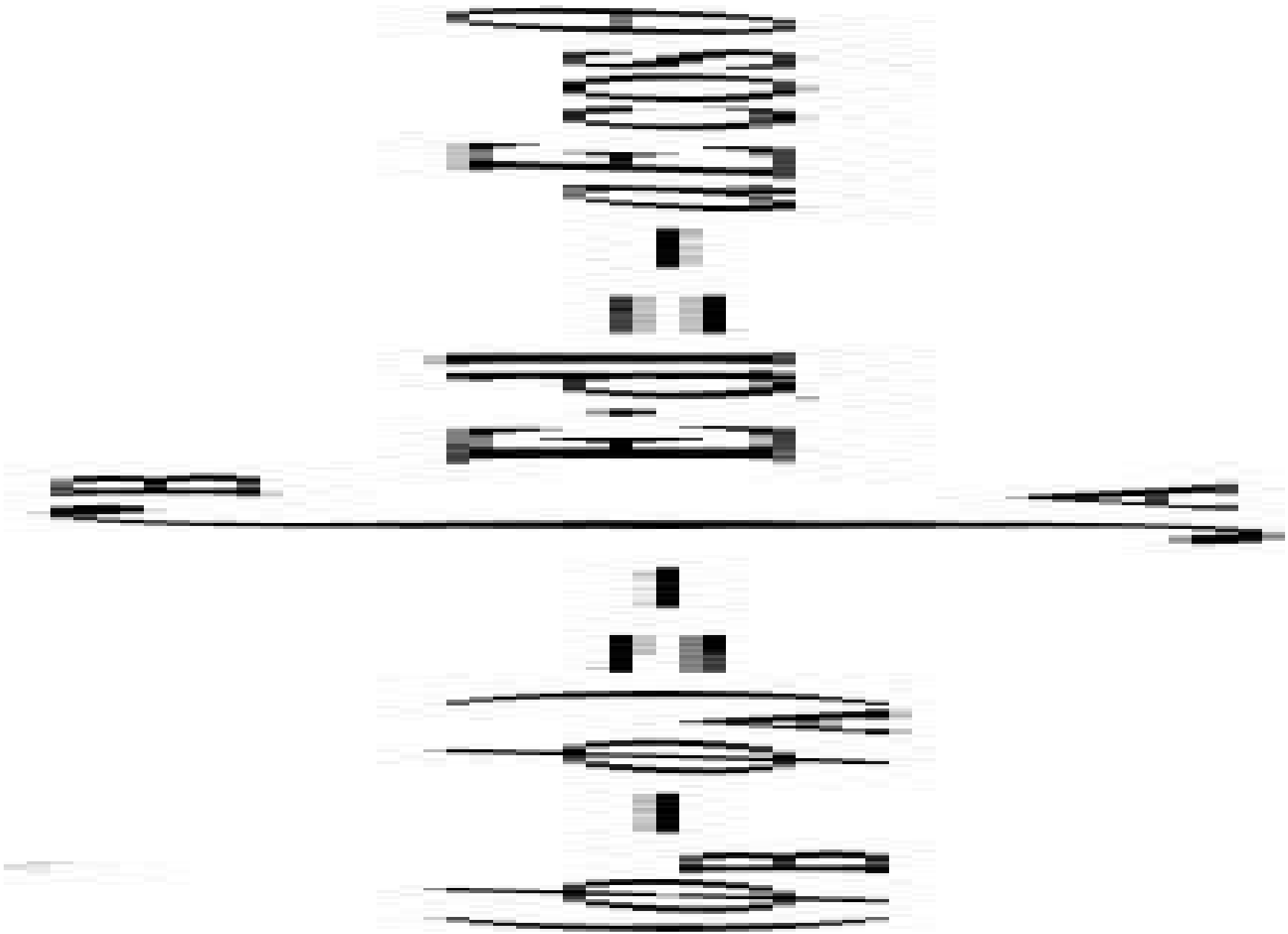
H atoms with net positive charges

The electron clouds of the sodium chloride and water molecules are sketched in Figure 2.6. Although the water molecule is not ionic, there is substantial electron transfer from hydrogen to oxygen, and the molecule has a permanent dipole moment almost as large as that of an alkali halide. Yet in the absence of an external electric field, there is no polarization in water, that is, the dipole moment per unit volume in a large-scale sample is zero. This is because thermal motion ensures the random orientation of the molecular dipole moments. However, water does become strongly polarized when an external field is switched on, and the relative permittivity of water at room temperature is 80*, much larger than the values listed in Table 2.1, which are all for non-polar molecules. The high relative permittivity occurs because the molecular dipole moments tend to line up along the direction of an external field. To see why this should be so, we shall first investigate the behaviour of a single isolated dipole.

The dipole in Figure 2.7 consists of charges $\pm q$ separated by a distance a. The dipole moment p is of magnitude $p = qa$, and points along the line joining the charges, at an angle θ to a uniform external field E. There is no net force on the dipole, but a torque $qaE \sin \theta$ acts in the sense required to line up the dipole parallel to the field. Work must be done against this torque to turn the dipole, whose potential energy therefore depends on its orientation. The potential energy of the dipole is



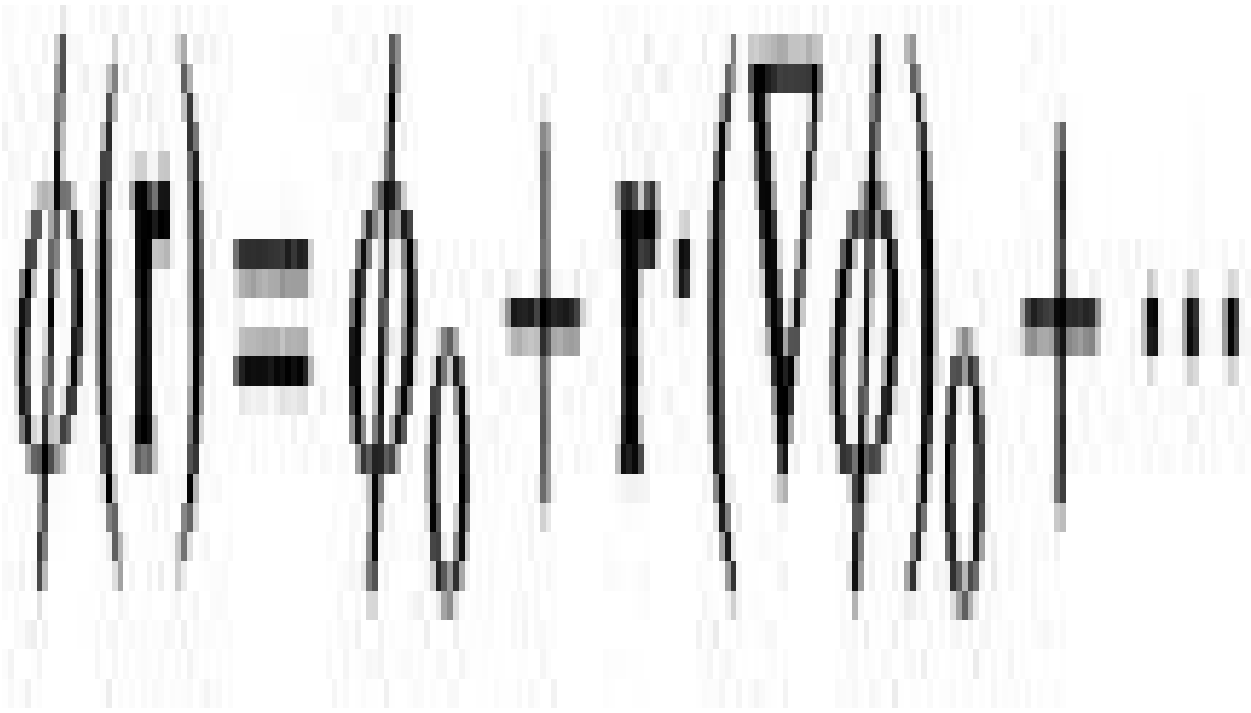
where ϕB and ϕA are the potential of the field E at the positions of + q and – q respectively. (There should strictly be an additional term to allow for the potential energy of the two charges in one another's fields. But for fixed separation this energy is constant, and we don't have to worry about it.) The potential difference (ϕB — ϕA) is given by



and

(2.11)

This result holds for polar molecules as well as for the idealized dipole made up of point charges. Taking the origin to be at the centre of a particular molecule, the potential $\phi(r)$ can be expanded in a Taylor series as



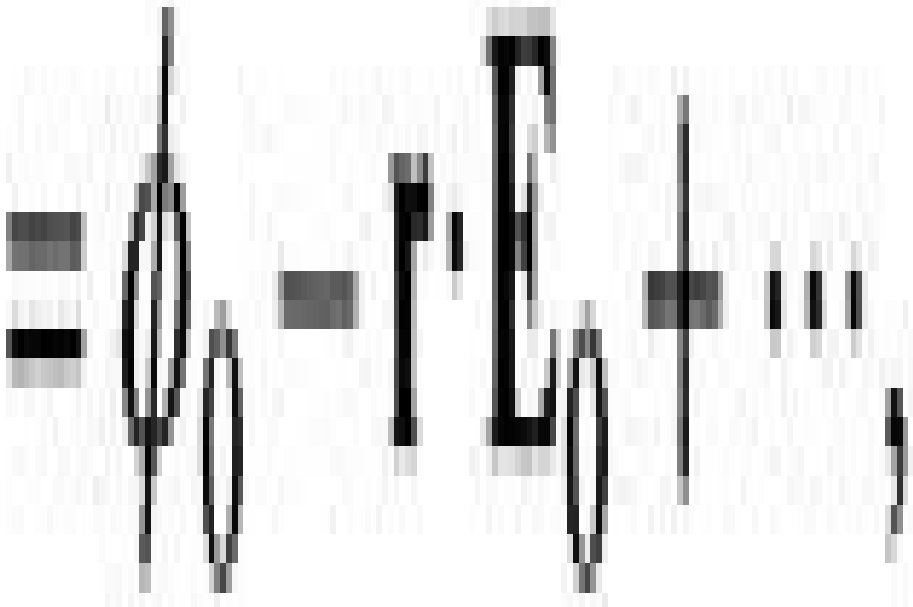
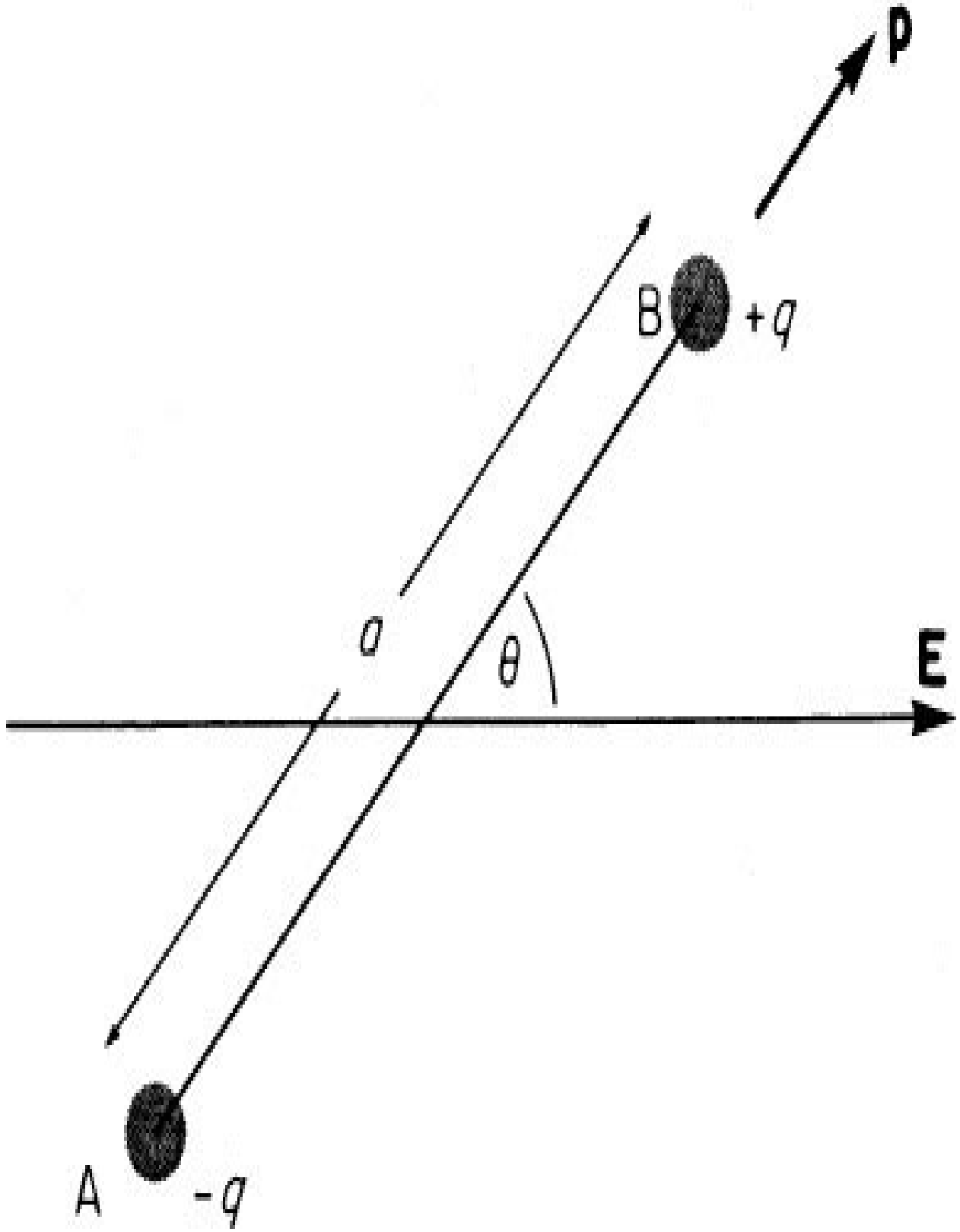
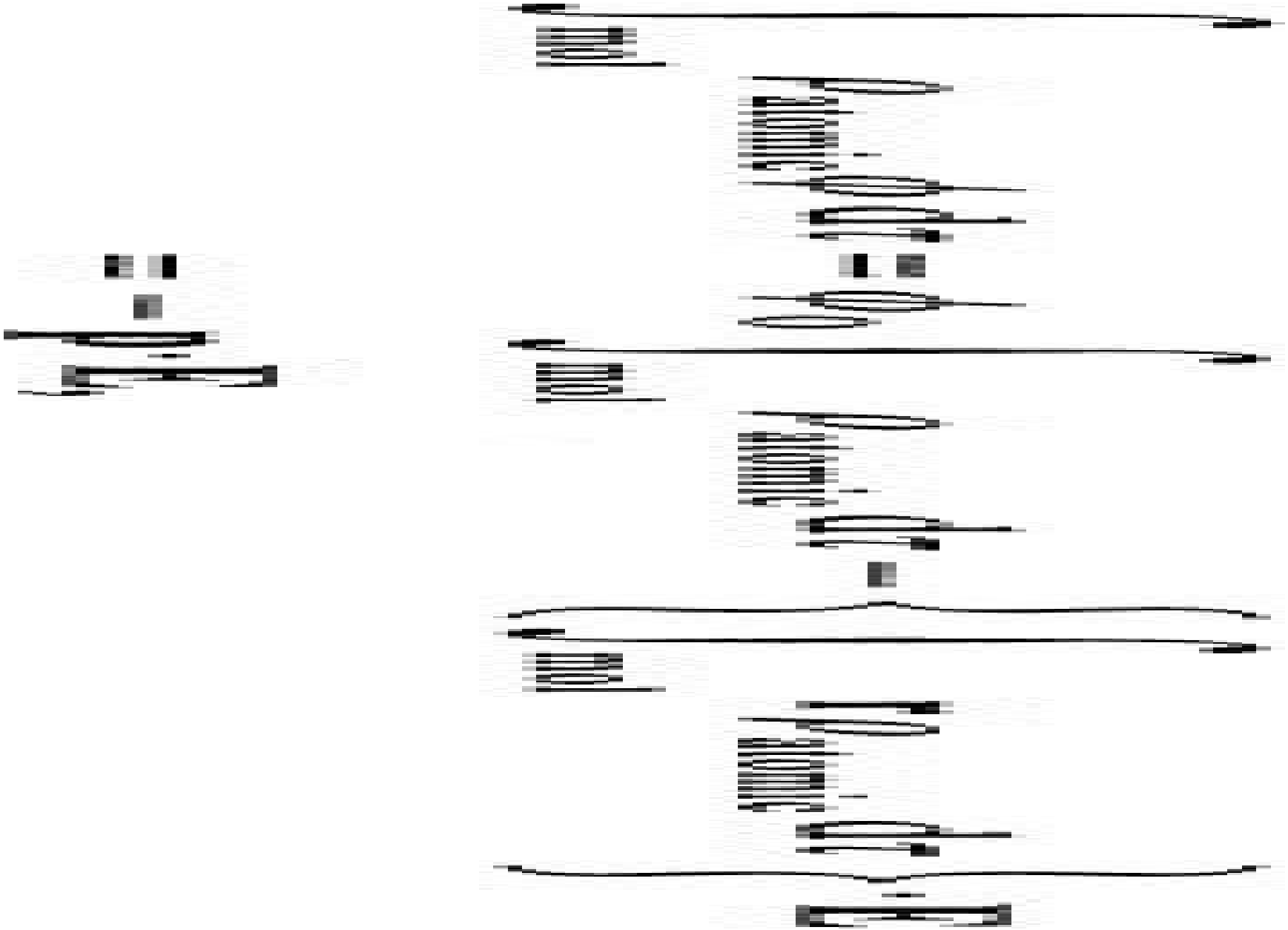


Figure 2.7. A dipole in an external field.


P
B
+q
a
θ
E
A
-q

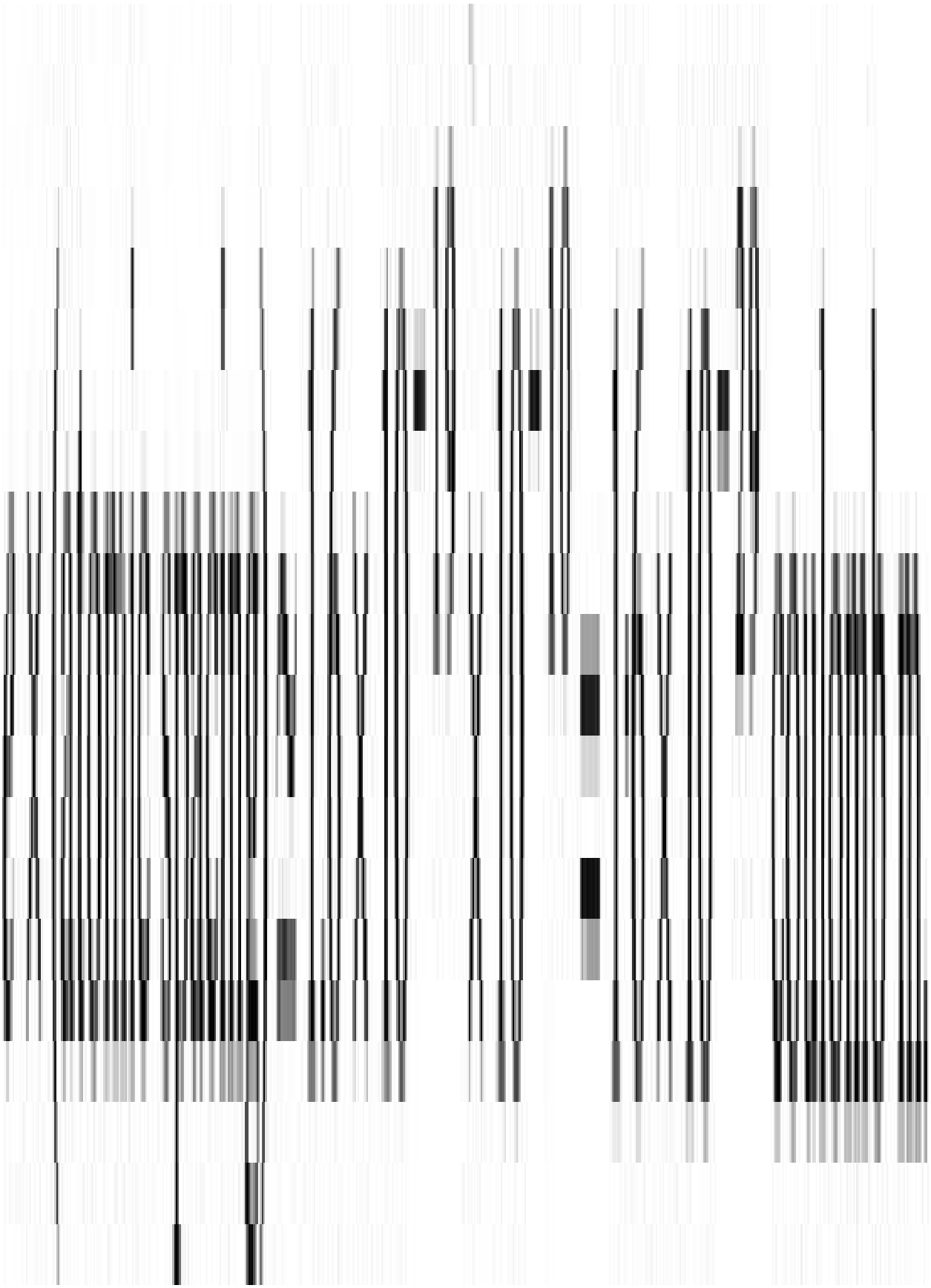
where ϕoand E0 are the potential and field at the origin. The potential does not change much over a distance comparable with the size of molecule, so the higher order terms are very small and may safely be neglected. The potential energy of the molecule then becomes



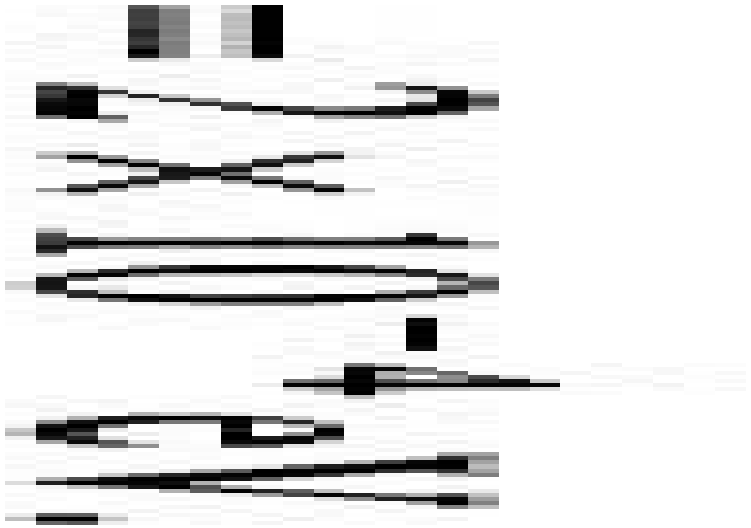
since ∫mol ρatomic dτ is zero for a neutral molecule.

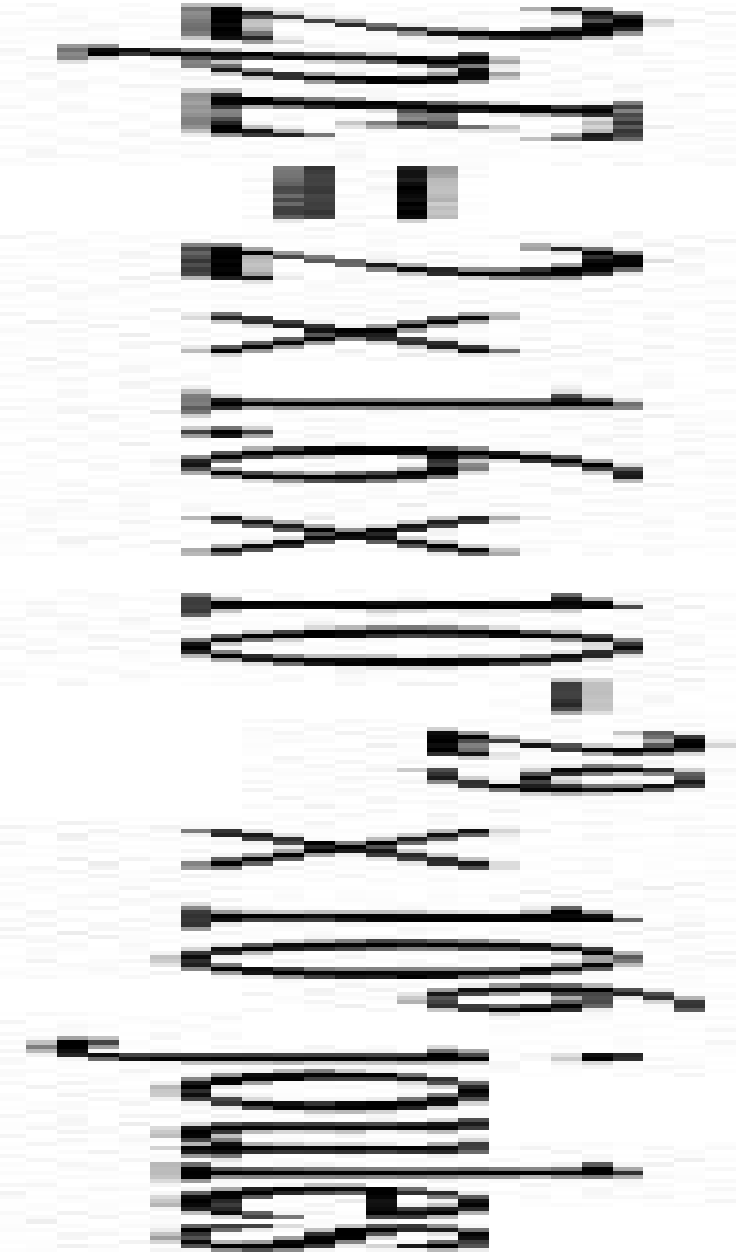
The susceptibility of a gaseous polar dielectric

The local field acting on a molecule in a gas is almost the same as the external field E. For a molecule with dipole moment p, Equation (2.11) shows how the potential energy of the molecule varies with its orientation. The energy has a minimum at cos θ = 1, when the dipole is aligned along the field. An amount of energy 2p E must be expended to reverse the dipole moment and point it in the direction opposite to the field. How much energy is this? If one electron in the molecule has been transferred from one atom to another, the resulting dipole moment is roughly



Typical polar molecules have dipole moments of this order of magnitude, and in a field E of 10^6 volts/m (about as big as can be achieved in a gas),

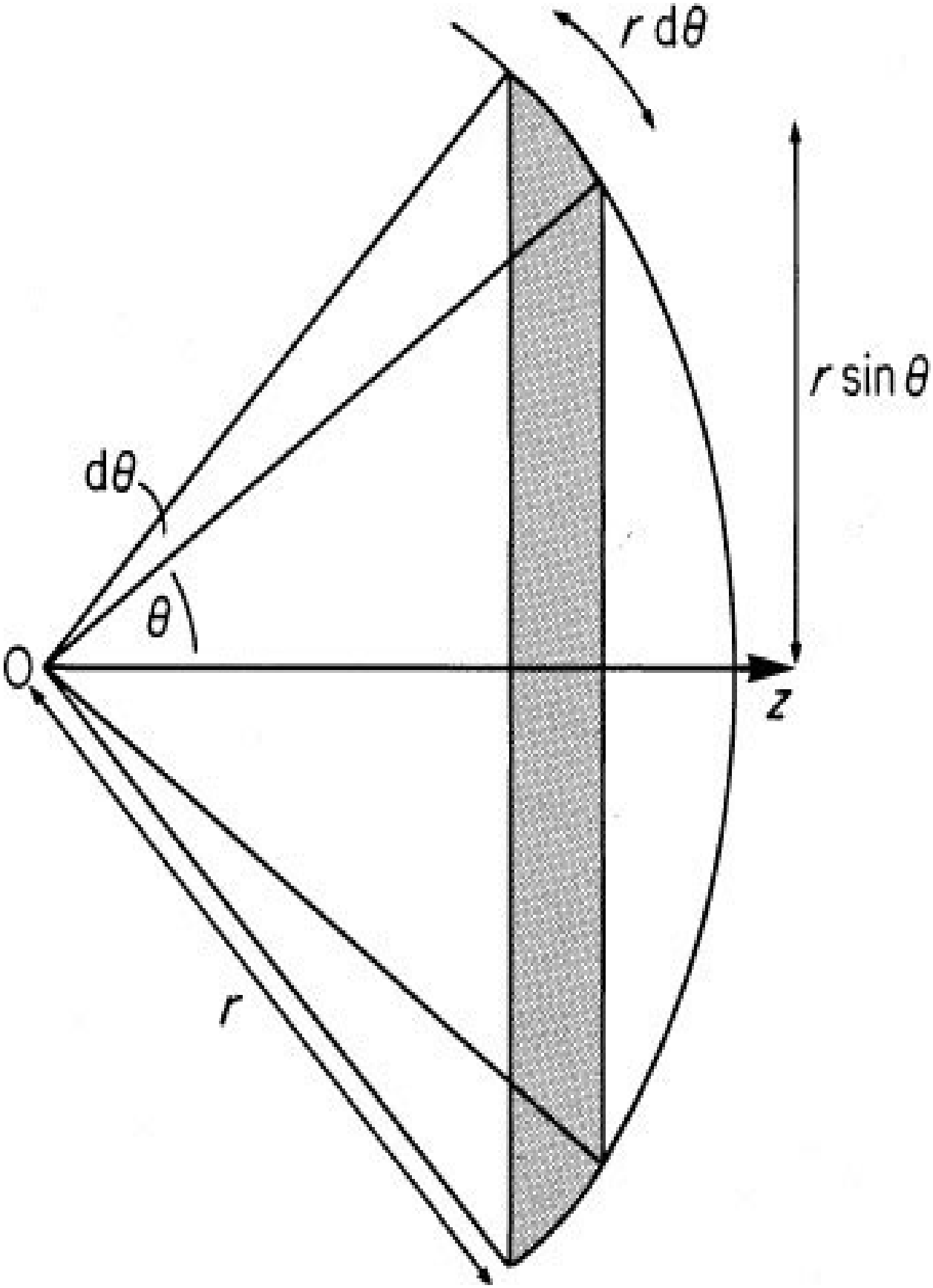


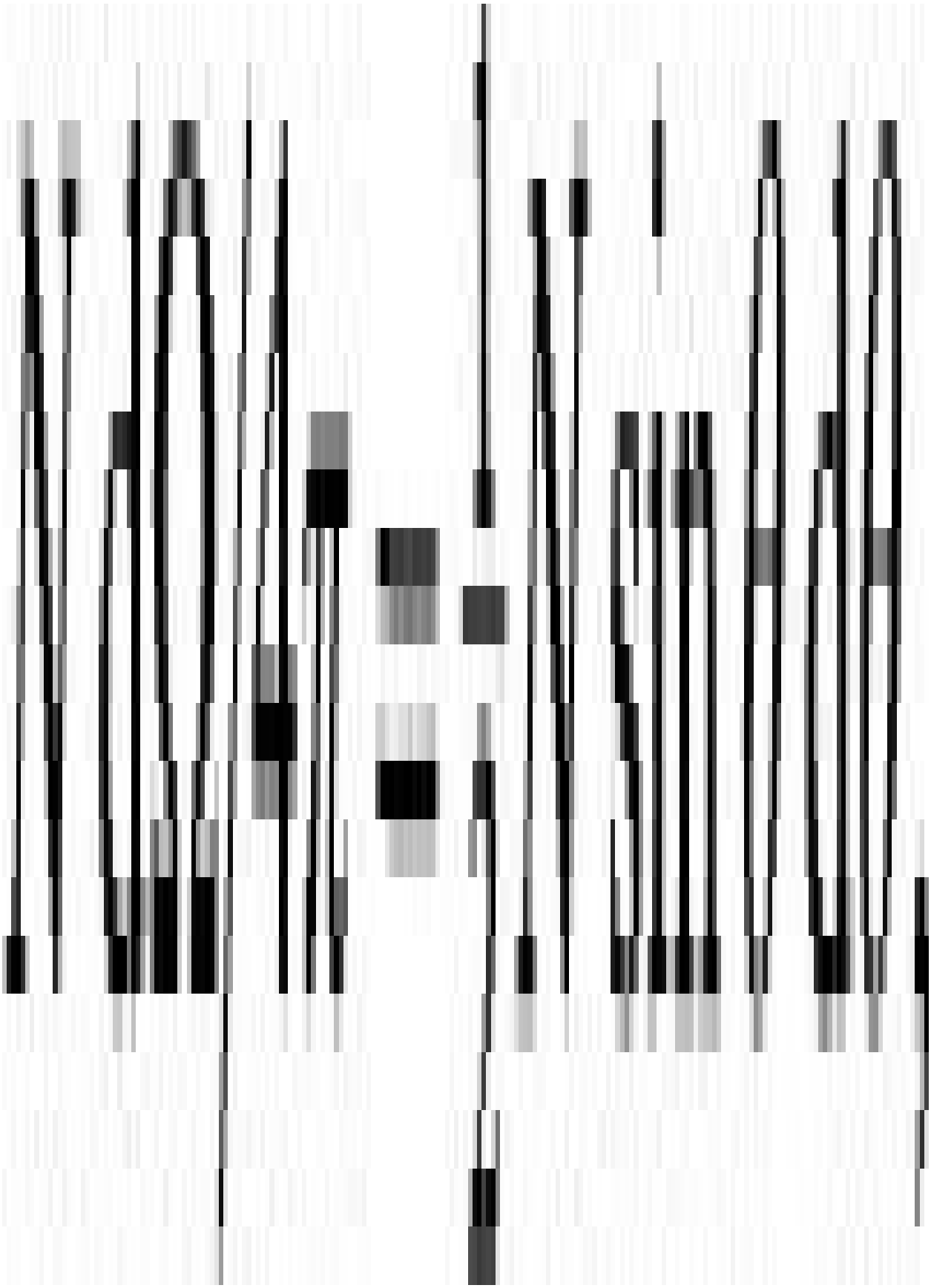


At ordinary temperatures this is much less than the energy associated with thermal motion. The average kinetic energy (where k is Boltzmann's constant and T the absolute temperature) of each molecule in a gas is about 0.04 eV at room temperature. The gas molecules are continually knocked about in collisions, changing their orientations and directions of motion, and transferring kinetic energy from one to another. Because the potential energy $-\mathbf{p}\cdot\mathbf{E}$ is so small compared with the kinetic energy, the random thermal motion is hardly affected by the presence of the field E, and molecules are still to be found in all orientations. However, there is a slight preponderance of molecules with dipoles lined up along the field, the direction in which their potential energy is a minimum. To calculate the net polarization caused by this alignment, we must know the probability of finding a molecule at any particular orientation to the field.

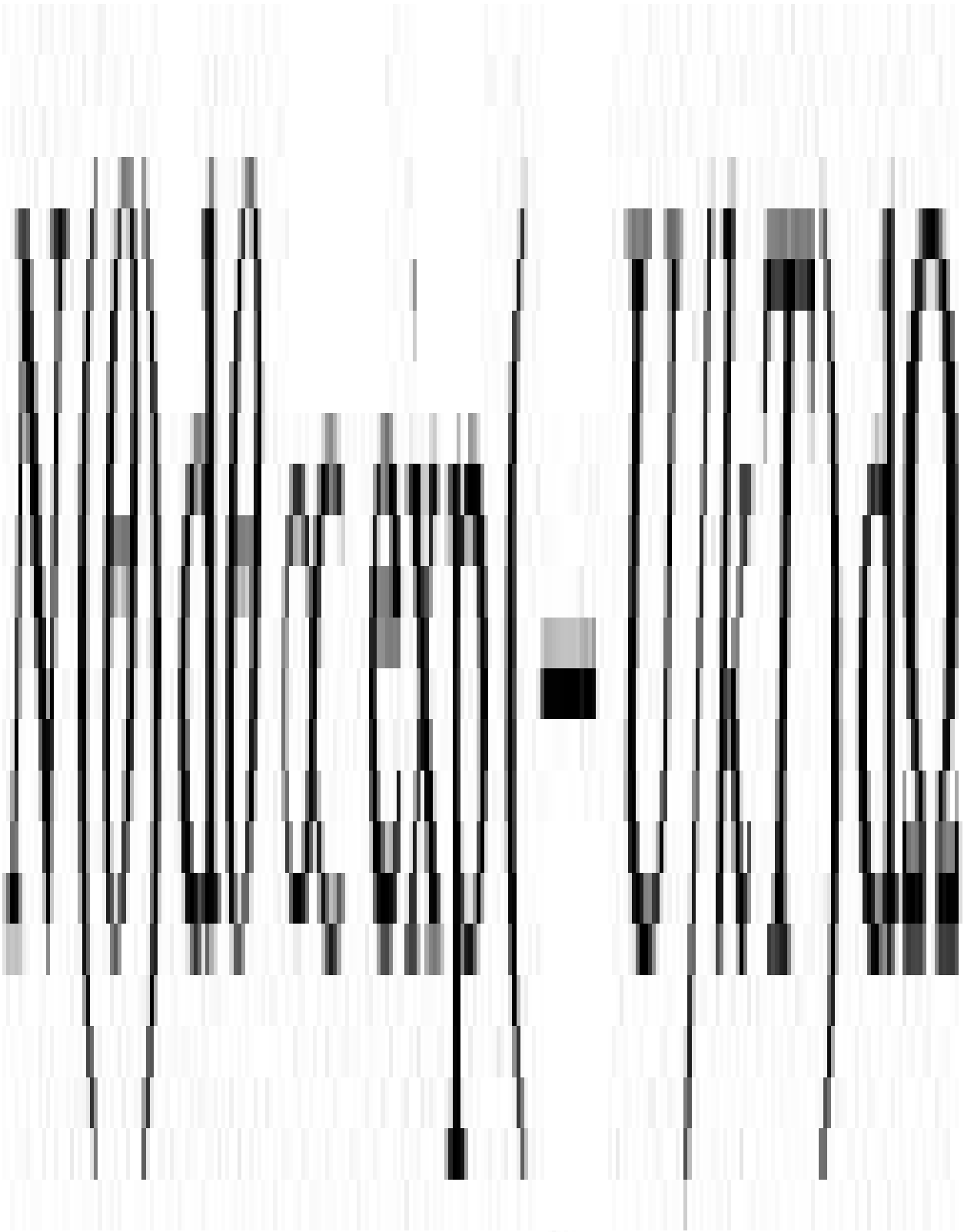
Before the electric field is switched on, all orientations of the molecules in the gaseous polar dielectric are equally probable. The number of molecules whose dipole moments have directions which lie within a solid angle $d\Omega$ is simply proportional to $d\Omega$. Now the solid angle included by the range of orientations in the cone between angles θ and $(\theta + d\theta)$ to the z-axis is $d\Omega = 2\pi \sin\theta\, d\theta$ (see Figure 2.8). When there is no electric field the density of molecules with dipole moments pointing in directions in the range $d\theta$ is

Figure 2.8. The shaded area represents the band between θ and $(\theta + d\theta)$ on the surface of a sphere of radius r. The area of the band is $dS = 2\pi r^2 \sin\theta\, d\theta$, and hence the solid angle $d\Omega = dS/r^2 = 2\pi \sin\theta\, d\theta$.

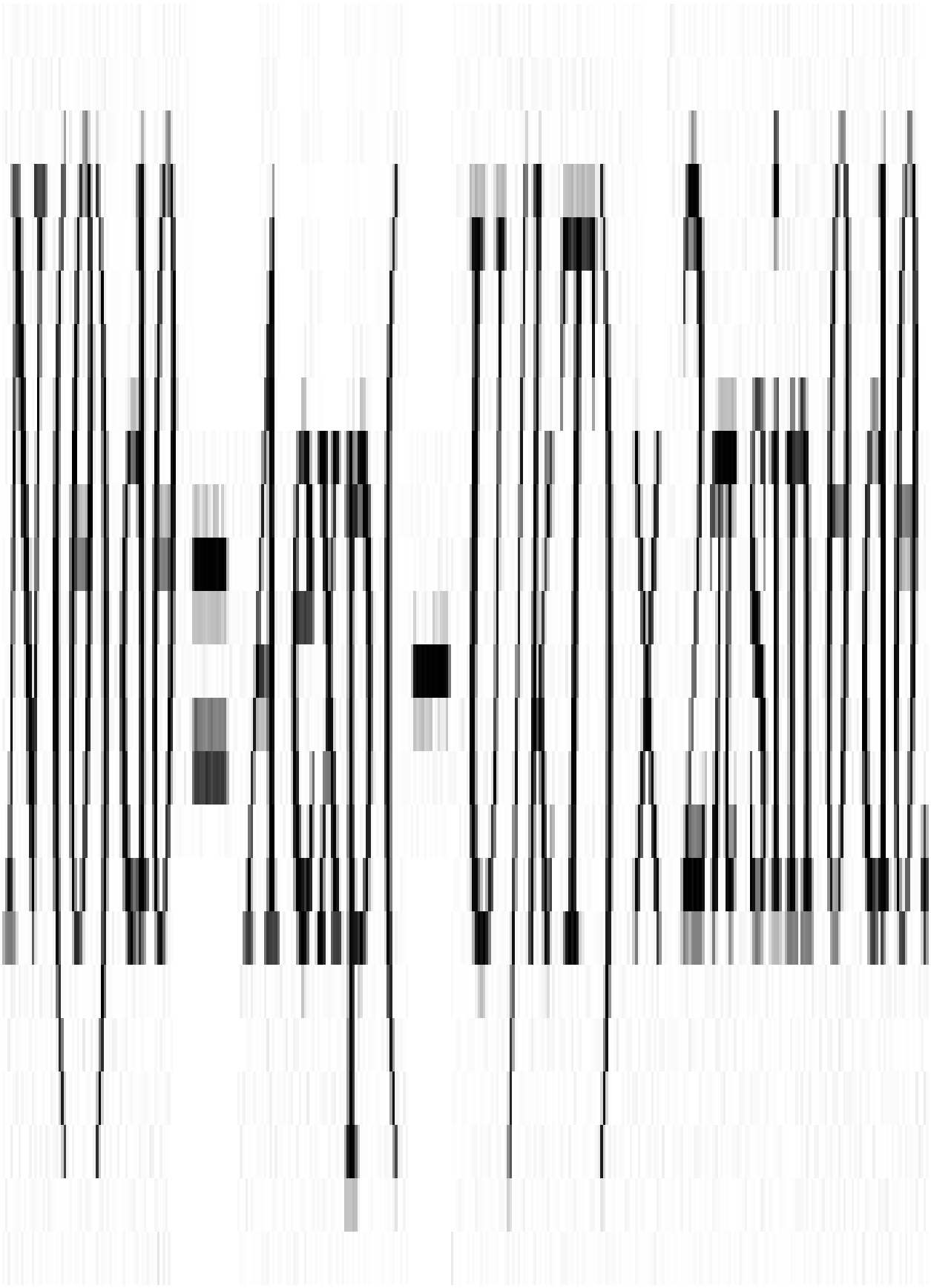

r dθ
r sin θ
dθ
θ
O
z
r



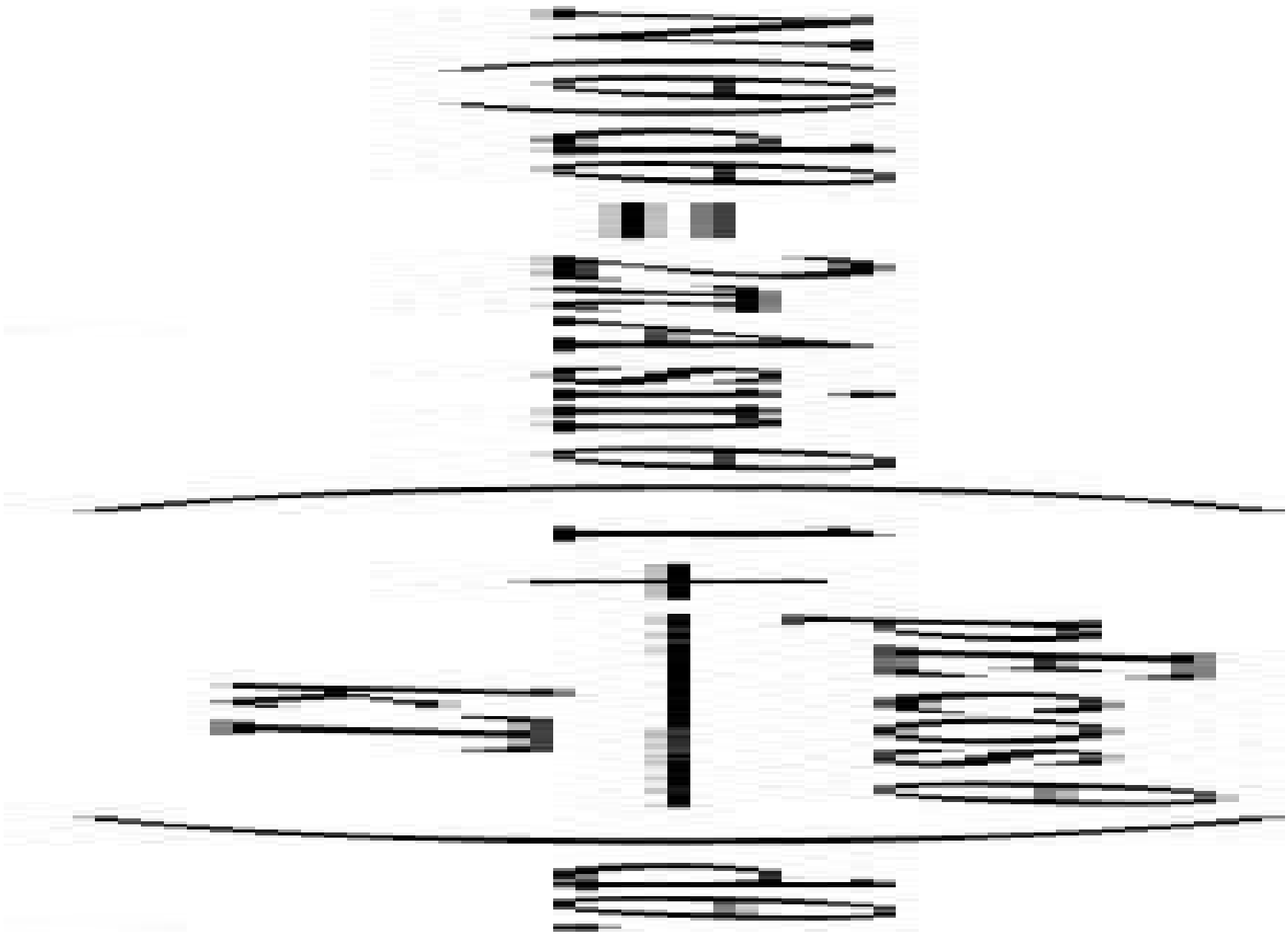
where N is the total number of molecules per unit volume. After switching on an electric field E directed along the z-axis, a molecule with dipole moment at θ acquires an additional potential energy $U = -pE\cos\theta$. The probability distribution at thermal equilibrium is now modified by the presence of a Boltzmann factor* $\exp(-U/kT)$. If in the presence of the field, the number of molecules per unit volume with dipole moments in the range $d\theta$ is $N(\theta)d\theta$, then



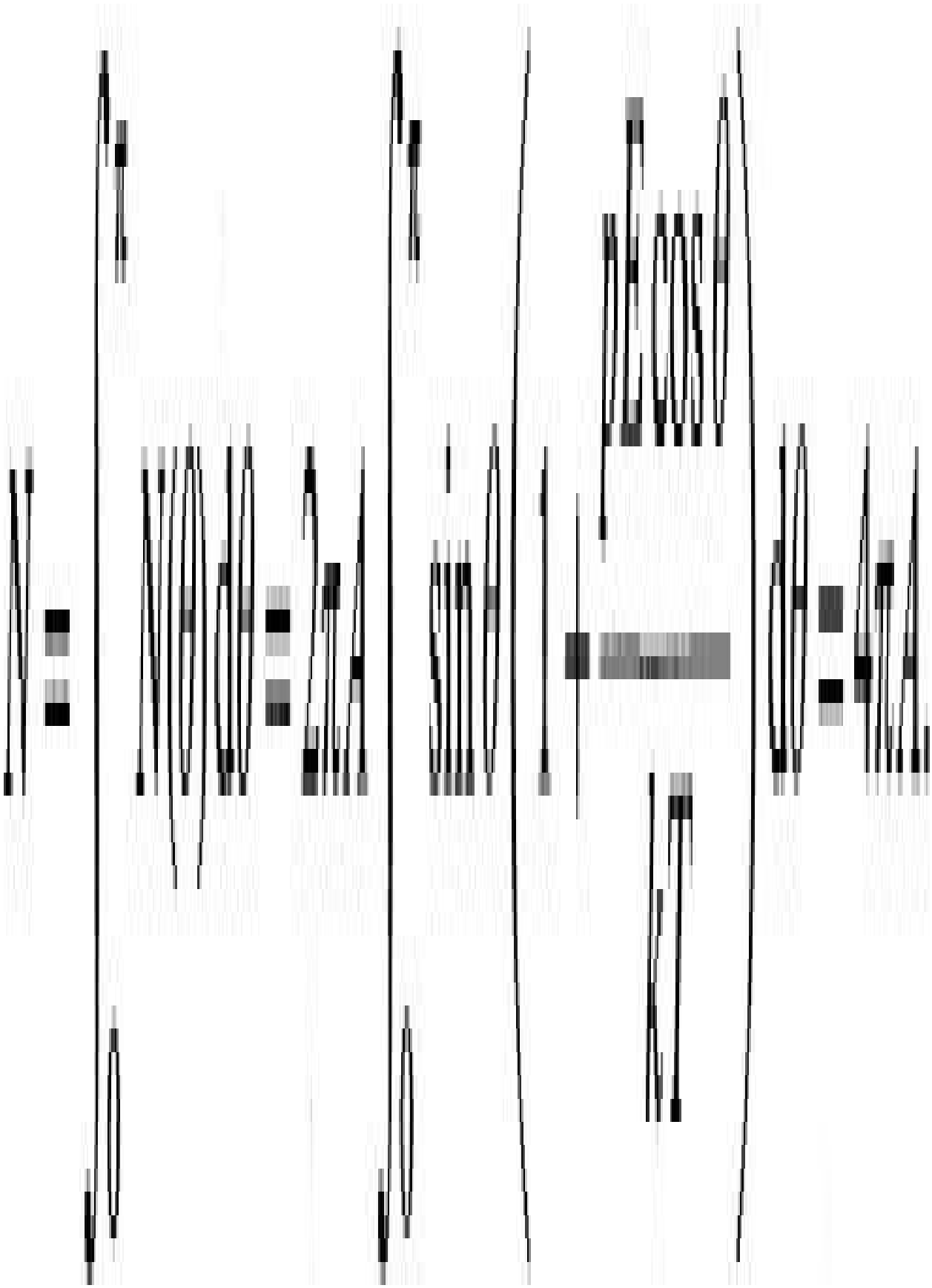
or



The constant of proportionality A is determined by the requirement that the total number of molecules per unit volume at all orientations is still N. Because U/kT is normally very small, the Boltzmann factor can be expanded in powers of U/kT with the retention of only the first term: exp (− U/kT) 1 − U/kT = + pE cos θ/kT, leading to

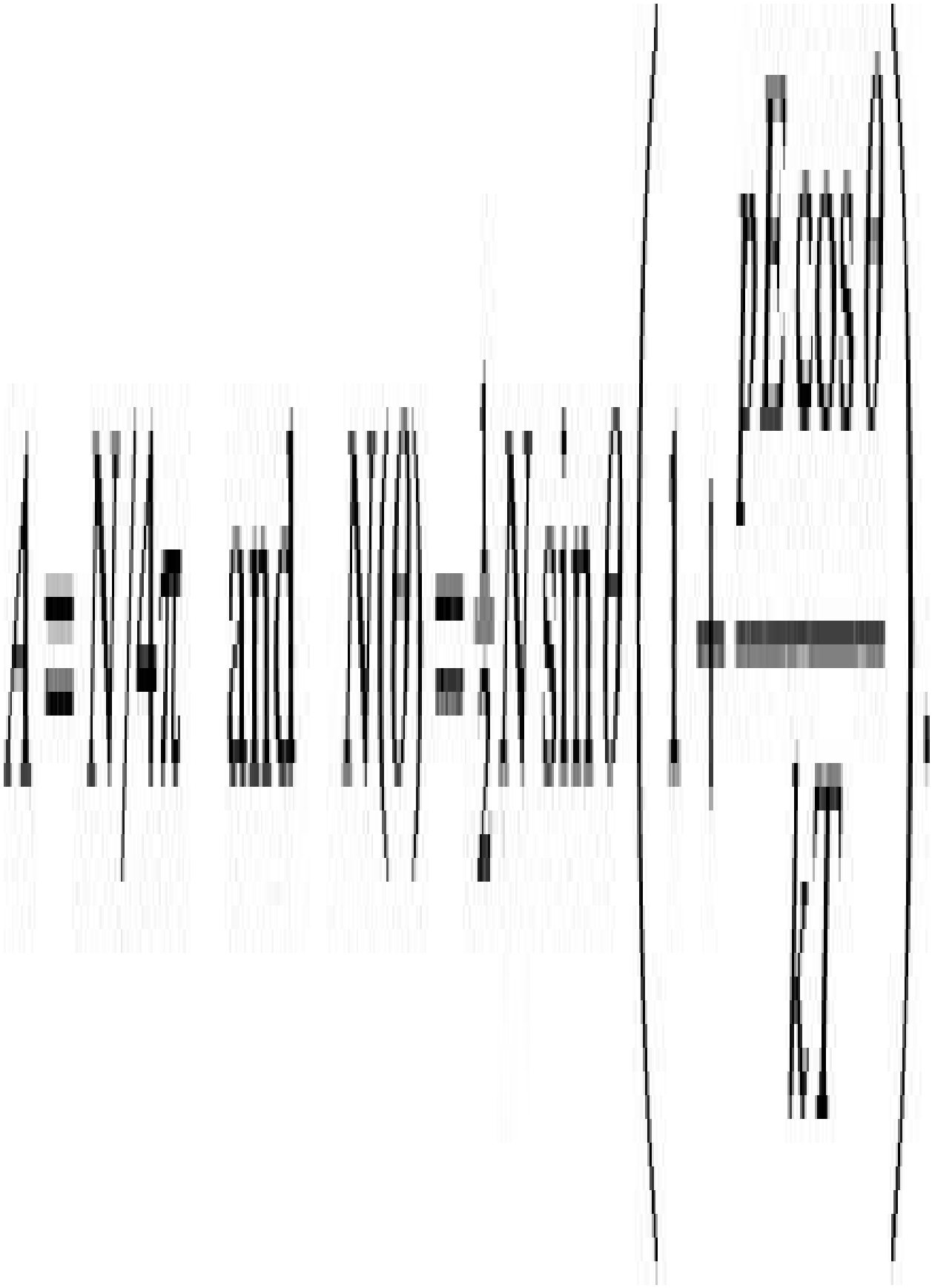


The total number of molecules per unit volume is



Hence

(2.12)



In Equation (2.12) the term in cos θ is small, and there are almost the same numbers of molecules parallel to the field (cos θ = 1) and opposite to the field (cos θ = − 1). Nevertheless, in a volume large enough to contain many molecules, there is a net dipole moment, which by symmetry must be in the direction of the field. For a dipole moment pointing in the direction θ, the component along the field is p cos θ. To find the net dipole moment per unit volume, we must add the components along the field for all the molecules, that is, we must integrate p cos θ over the probability distribution. The dipole moment per unit volume is

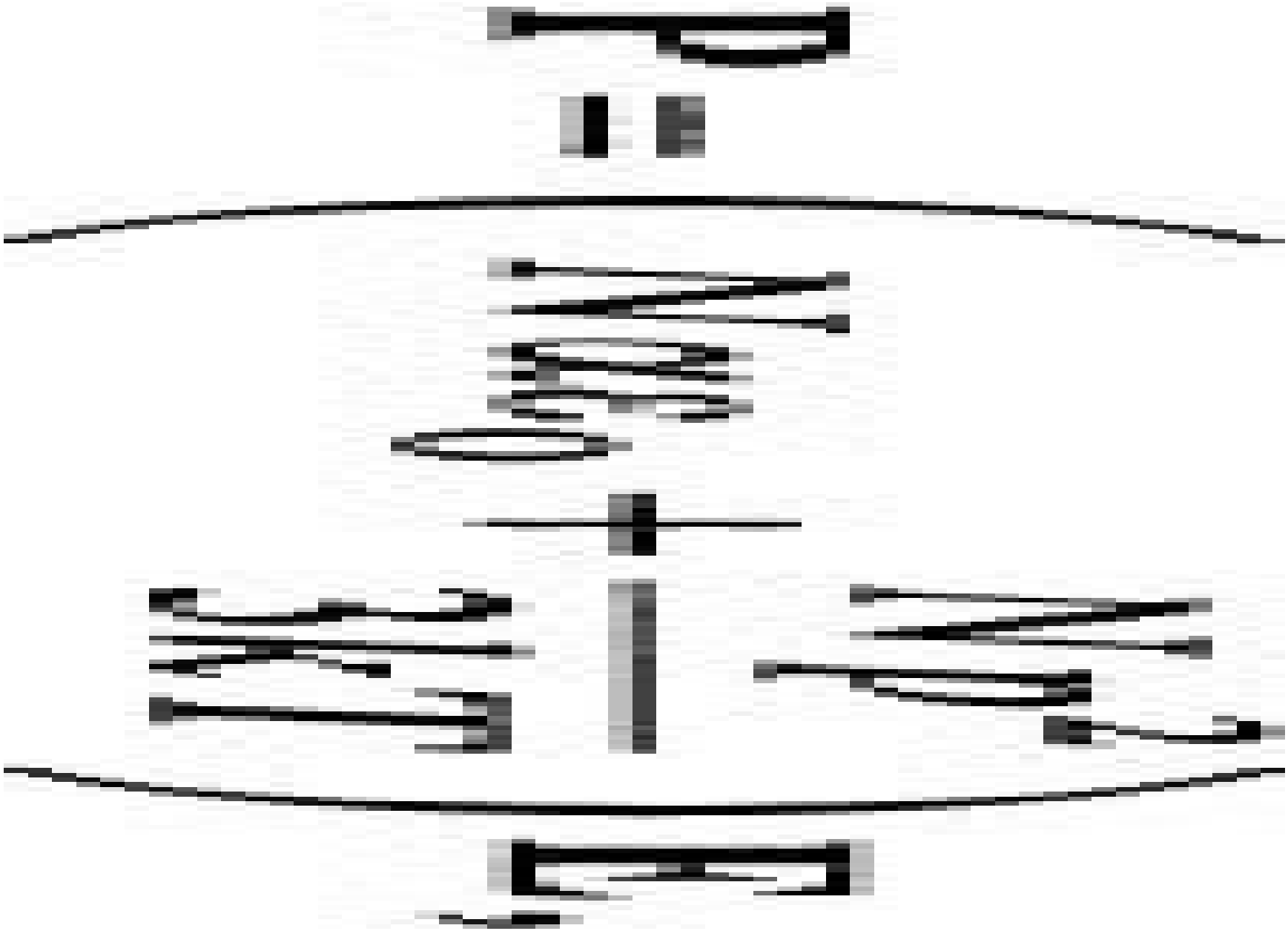
(2.13)

$$P = \int_0^{\pi} N(\theta) p \cos\theta \, d\theta$$

$$= \frac{N}{2} \int_0^{\pi} \sin\theta \left(1 + \frac{pE\cos\theta}{kT}\right) p \cos\theta \, d\theta$$

$$= \frac{Np^2E}{3kT}.$$

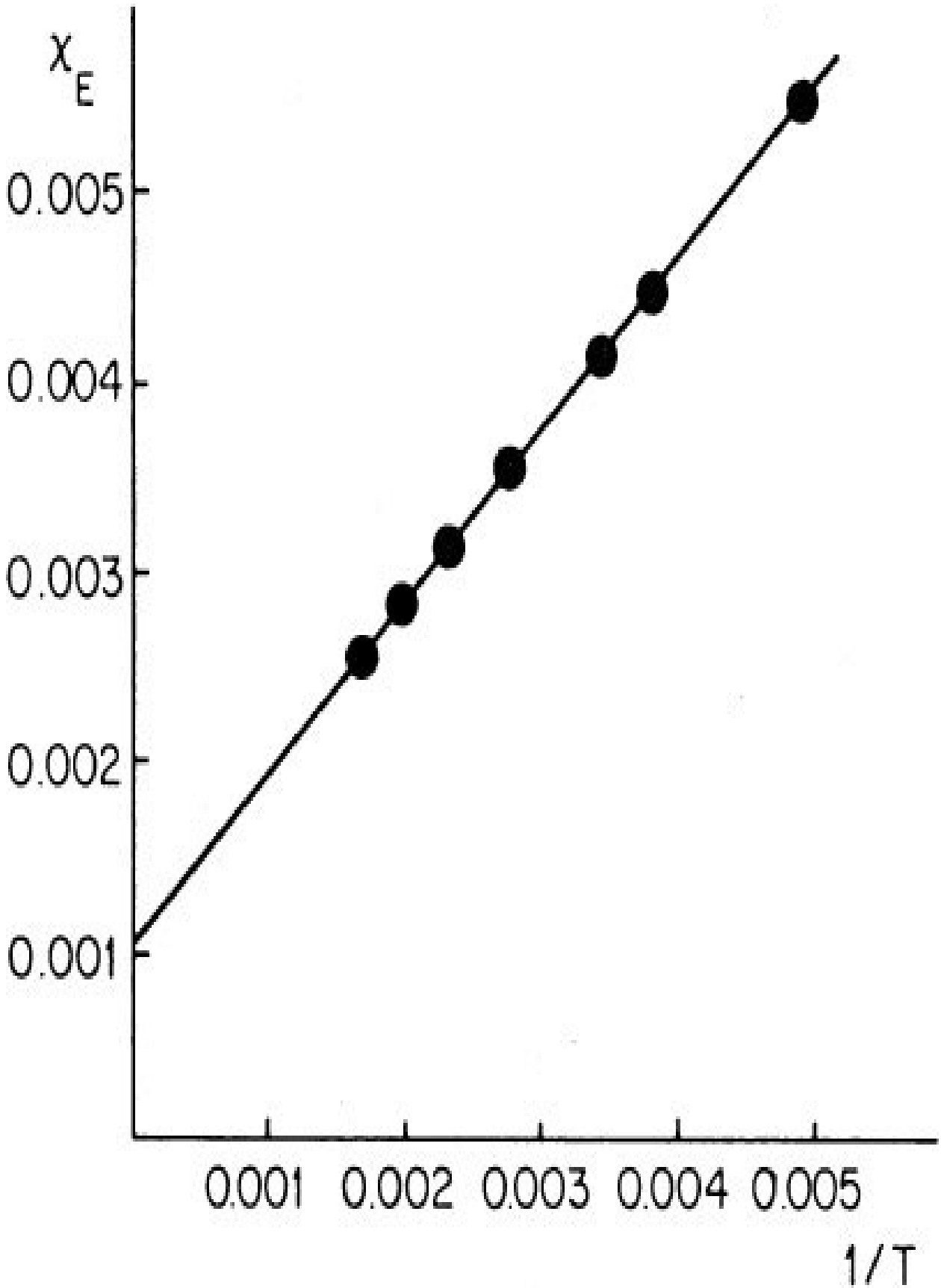
All the molecules also acquire an induced dipole moment, which points along the field whatever may be the direction of the permanent dipole moment. If the molecular polarizability is α, there is an additional induced polarization Nαε0E. Finally, we have the result that the net polarization of the gaseous polar dielectric is



and its electric susceptibility

(2.14)

Figure 2.9. The temperature dependence of the susceptibility of HCl gas at a density 1/22.4 mole/litre. (Data from Zahn, Phys. Rev. 24, 400 (1924).)


χ_E
0.005
0.004
0.003
0.002
0.001
0.001
0.002
0.003
0.004
0.005
1/T

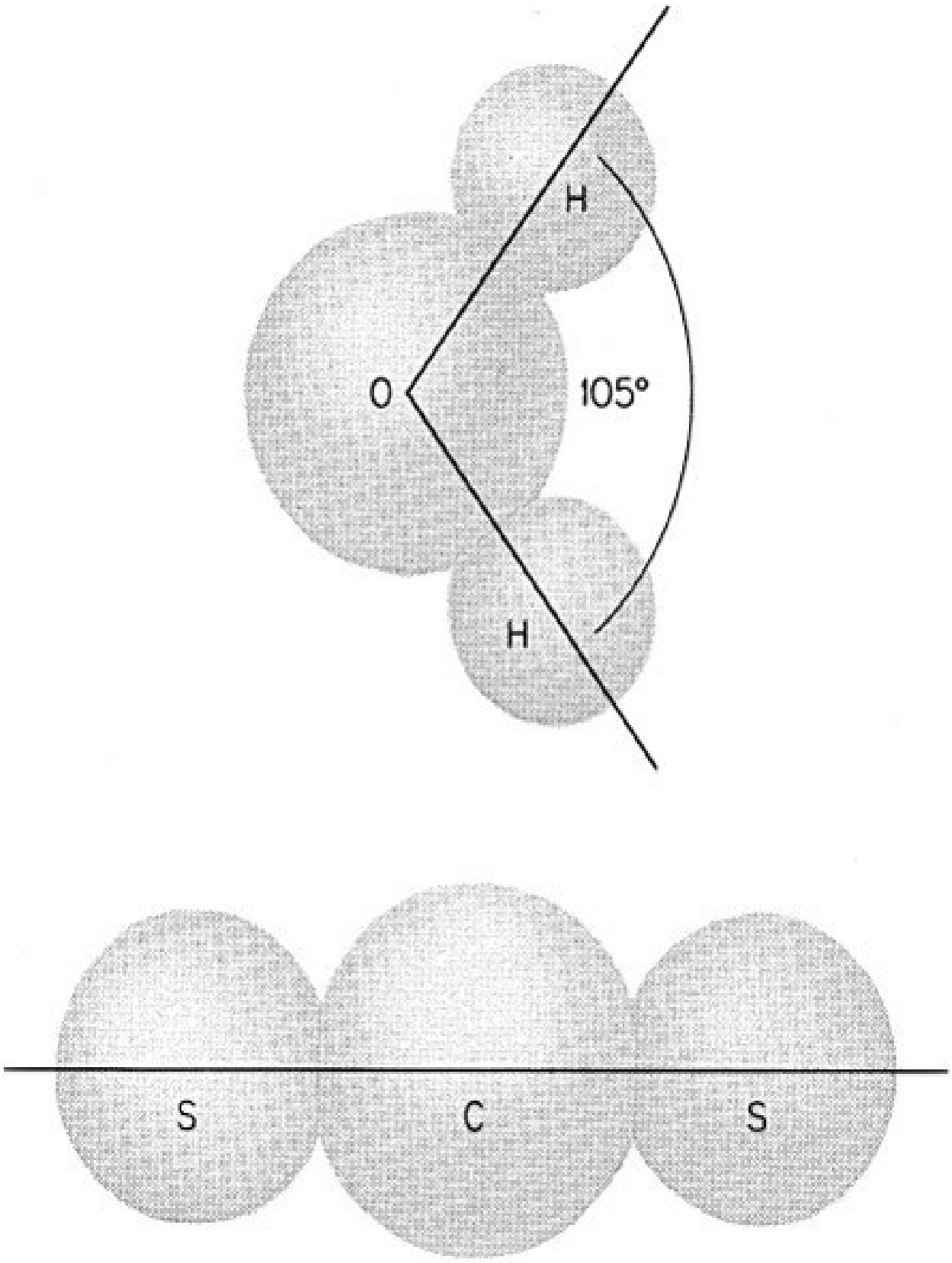
By measuring susceptibilities at different temperatures, it is possible to distinguish between permanent and induced dipole moments. In Figure 2.9, χE is plotted against 1/T for HCl gas at a fixed density of 1/22.4 moles/litre (i.e. the density found at S.T.P.). As expected from Equation (2.14), the plot is a straight line. The intercept at $1/T = 0$ represents the susceptibility due to induced polarization at this density, while from the slope the permanent dipole moment can be deduced to be 3.6×10^{-30} coulomb metres. To achieve this dipole moment, charges $\pm e$ would need to be separated by a little more than 2×10^{-11} m, or about one-fifth of a typical atomic diameter.

The susceptibilities and dipole moments of a number of gases are listed in Table 2.2. It is interesting to observe that carbon disulphide has no dipole moment although, like water, it contains two identical atoms bound to a common partner. There is electron transfer across the C—S bonds in carbon disulphide, and dipole moments associated with each bond. But the atoms in the CS2 molecule lie on a straight line, and the dipole moments exactly cancel one another, whereas in water the two O—H bonds make an angle of 105° (Figure 2.10). By glancing at a table of electric susceptibilities one can learn something about the shape of molecules!

Table 2.2. Dipole moments of polar molecules.

SUBSTANCE	SUSCEPTIBILITY OF GAS AT S.T.P.	DIPOLE MOMENTS IN
HCl	0.0046	3.6×10^{-30}
Ethane (C2H6)	0.0015	0
Ethyl alcohol	0.0061	5.7×10^{-30}
H2O	0.0126	6.2×10^{-30}
CS2	0.0029	0

Figure 2.10. The water molecule has a permanent dipole moment, but the linear CS2 molecule has none.


H
O
105°
H
S
C
S

2.2.3 Non-polar liquids

The molecules in a solid or liquid dielectric are so close together that they are affected by one another's short-range fields, and the average field Elocal acting on an individual molecule is not the same as the macroscopic field E. However, if the molecules do not interact chemically, they have the same shape as in the gaseous state. The polarizability α of the molecules is unchanged, since it depends only on their internal structure, and each one acquires an average dipole moment αε0Elocal. If the molecular density is N, the dipole moment per unit volume, i.e. the polarization, is

(2.15)

Generally it is difficult to estimate Elocal for the fixed molecules in solids. Liquids are more manageable, because the short-range fields may have a small average effect as neighbouring molecules continually change their relative positions and orientations. Even in a liquid, dielectric behaviour is complicated when polar molecules are present, since very large fields are generated by the permanent dipole moments. (Estimate the field at a distance of a few atomic diameters away from a water molecule—it is far, far bigger than any macroscopic field!)

But in a non-polar liquid the short-range fields are not so strong, and it is not too bad an approximation to assume that an individual molecule is surrounded by a continuous medium in which atomic structure has been smeared out. Like the real liquid, the continuous medium is polarized with a dipole moment per unit volume P so that polarization charges appear on any surface of the medium. Imagine that the molecule in question is removed, leaving a cavity embedded in the uniformly polarized medium. The field in the cavity is the sum of the

macroscopic field E and the field generated by polarization charges on its surface—and this sum is the field Elocal which would act on a molecule filling the cavity. Let us assume that the cavity is a sphere of radius R, and as in Figure 2.11 take the z-axis to be in the direction of E. At the point A in the figure, where the normal to the surface of the cavity is at an angle θ to E, the polarization charge on an element of surface dS is $\sigma_p(\theta)\, dS$, where

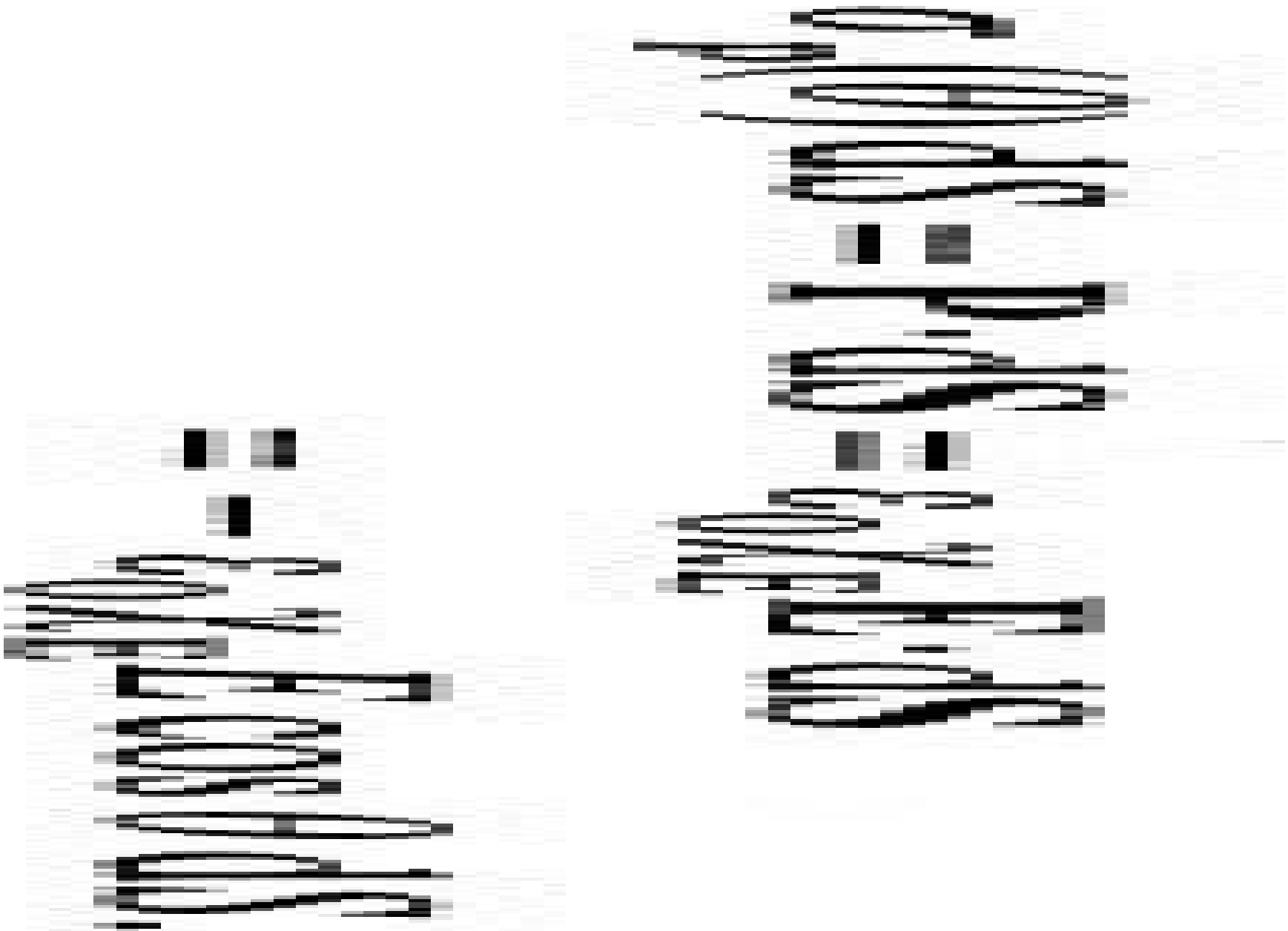
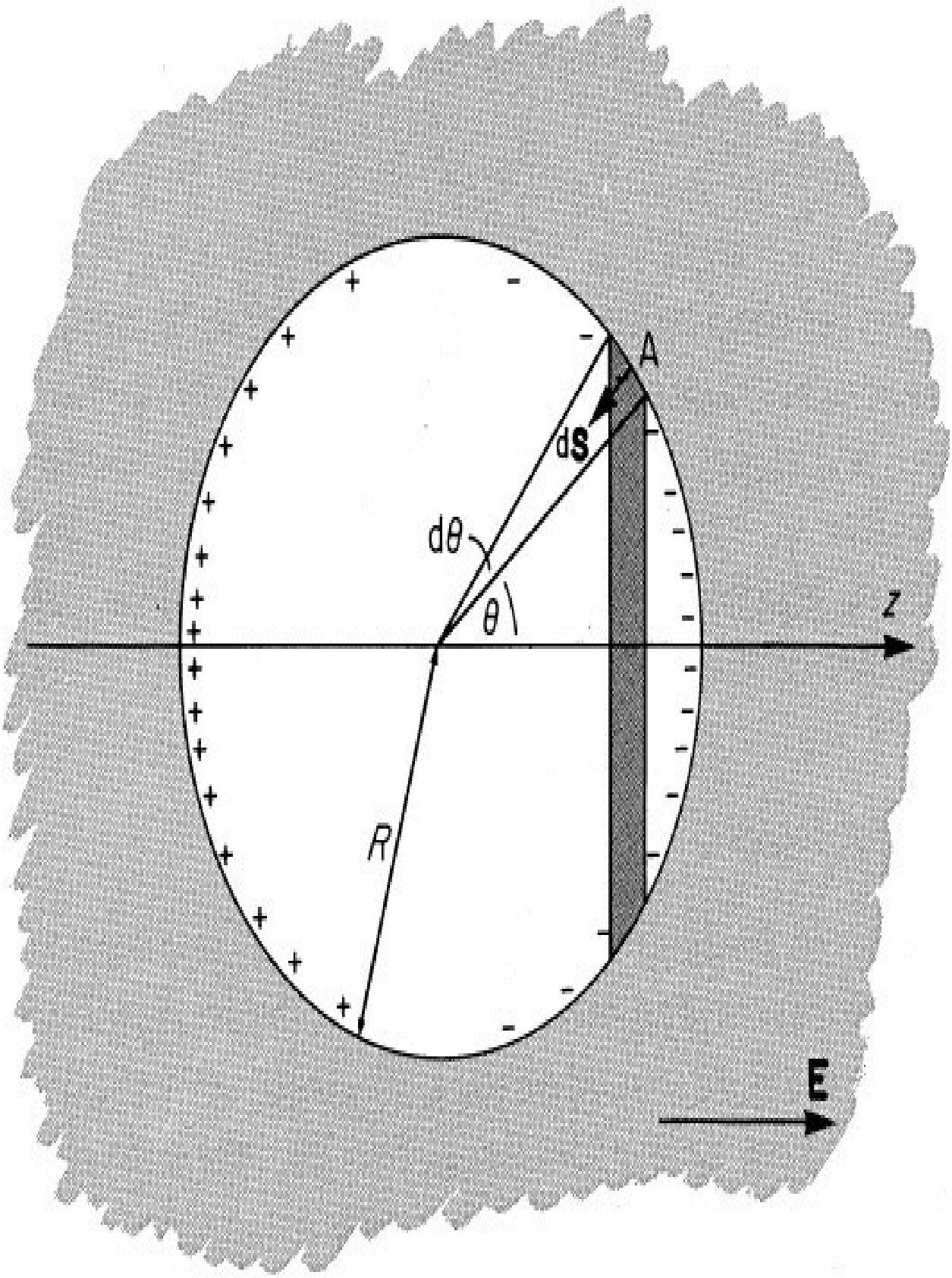
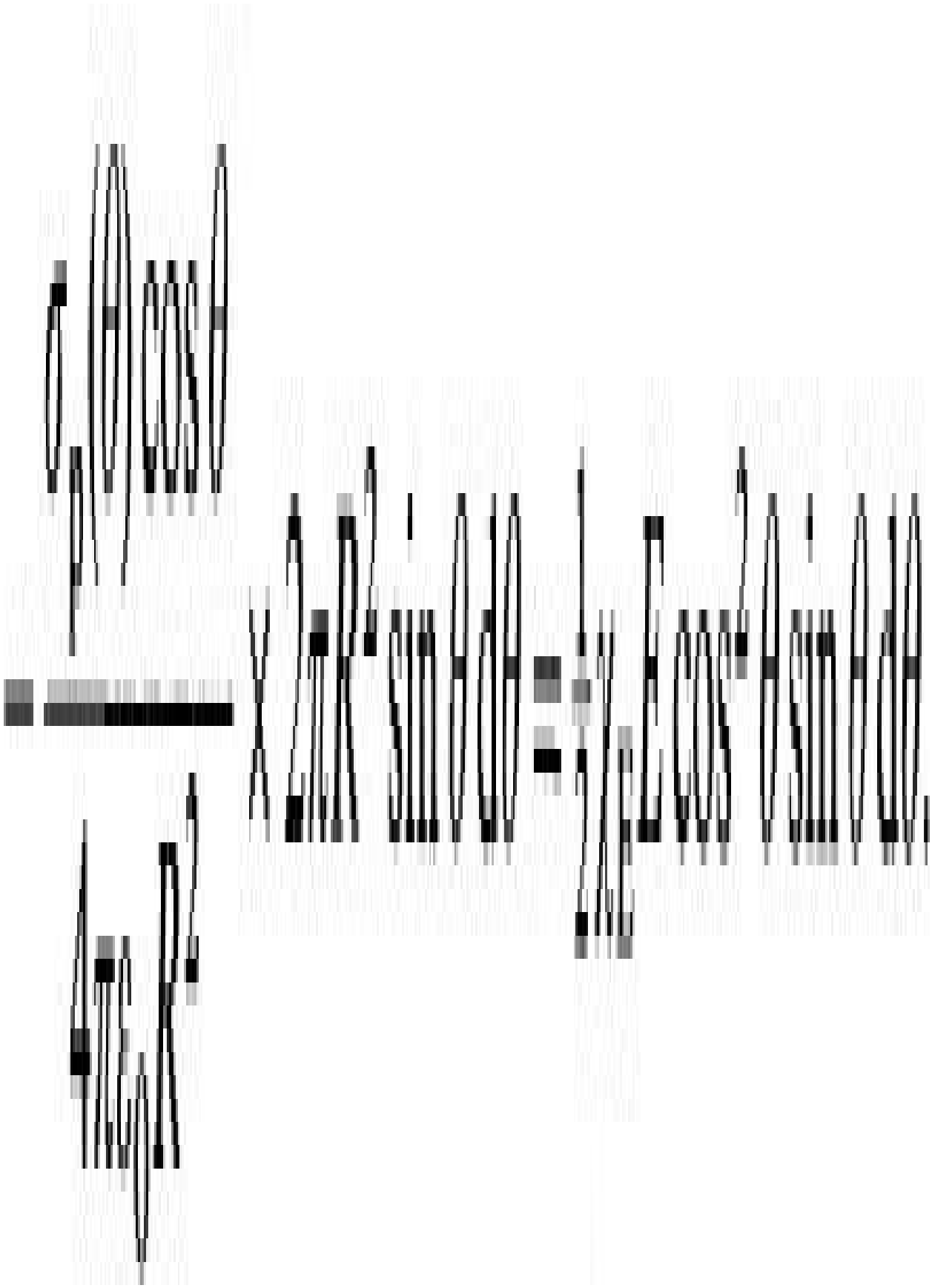


Figure 2.11. The field in a spherical cavity in a uniformly polarized liquid.

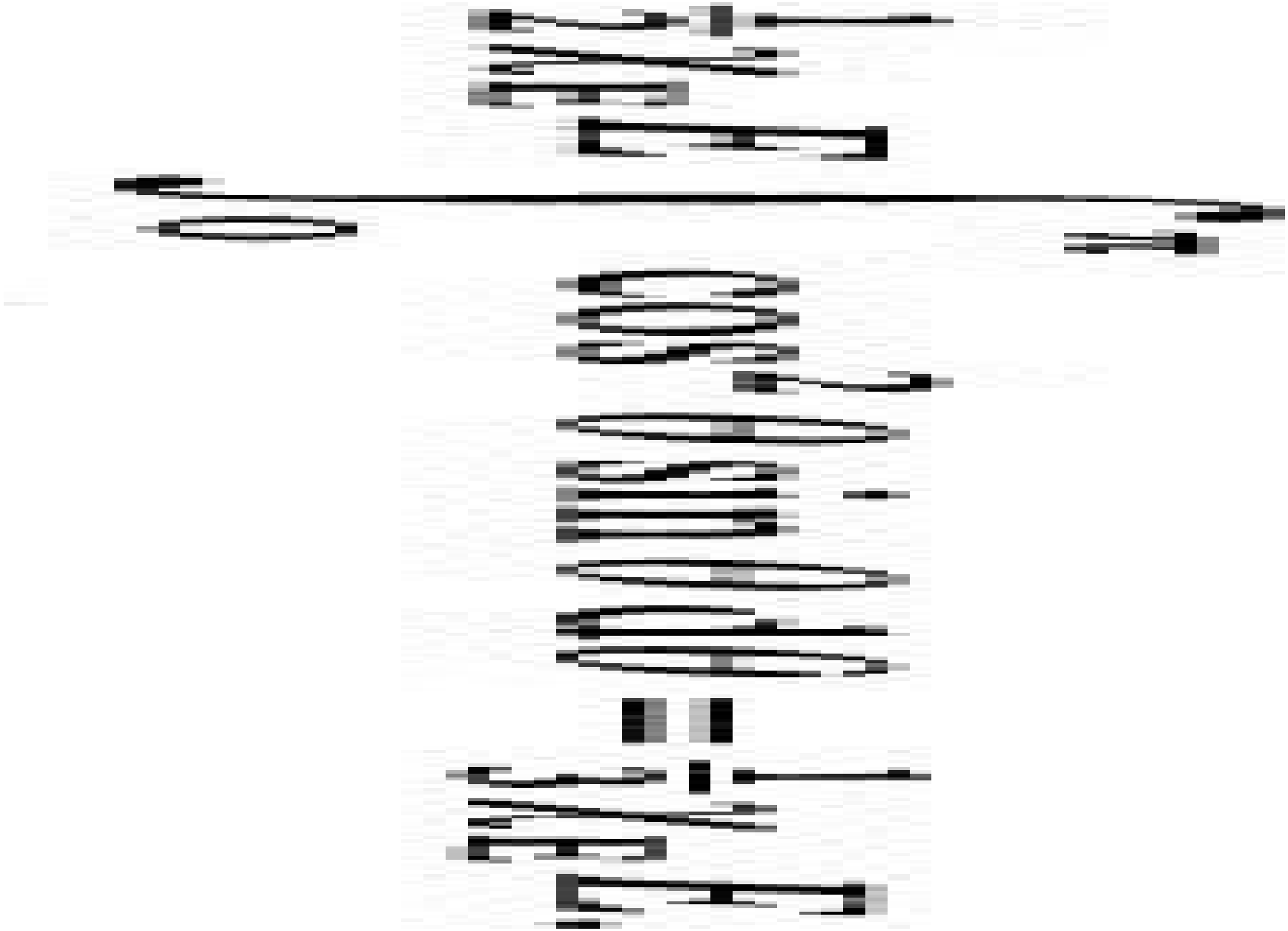

A
dS
dθ
θ
z
R
E

The minus sign arises because the outward normal to the dielectric is the inward normal to the cavity. By symmetry the net field generated at the centre of the cavity by the polarization charges is directed along the z-axis. At the centre the z-component of the field due to the charge $\sigma p(\theta)\, dS$ is

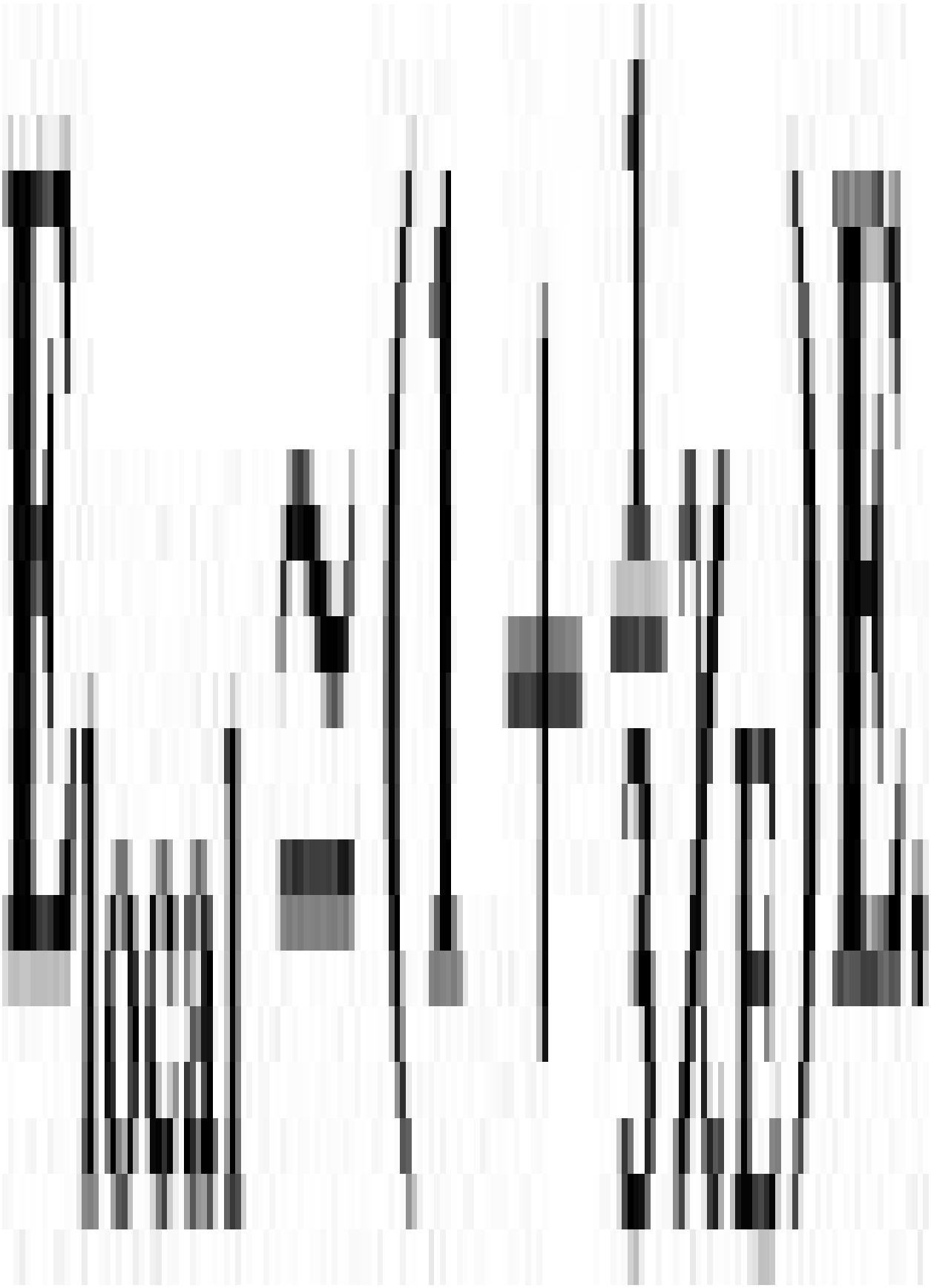
The area of the band on the surface of the cavity lying between θ and $(\theta + d\theta)$ is $2\pi R^2 \sin\theta \, d\theta$, and hence the field due to the band is



The field at the centre of the cavity generated by all the polarization charge is thus in the same direction as the macroscopic field E and is of magnitude



The field in the cavity is uniform (this is proved in section 3.3.2 of the next chapter), and so our estimate of the local field is



Substituting this value for Elocal in Equation (2.15) we find

Comparing this equation with the definition of susceptibility given in Equation (2.6), we see that

(2.16)

This approximation, which is known as the Clausius–Mossotti formula, cannot be expected to be very accurate since it is based on such a crude representation of the neighbouring molecules. However, the Clausius–Mossotti formula does demonstrate that polarization in the liquid dielectric causes the local field to be larger than the macroscopic field, and makes a reasonable estimate of the susceptibility in terms of molecular polarizability.

The molecular polarizability can be derived from the susceptibility of a gas (Equation (2.10)):

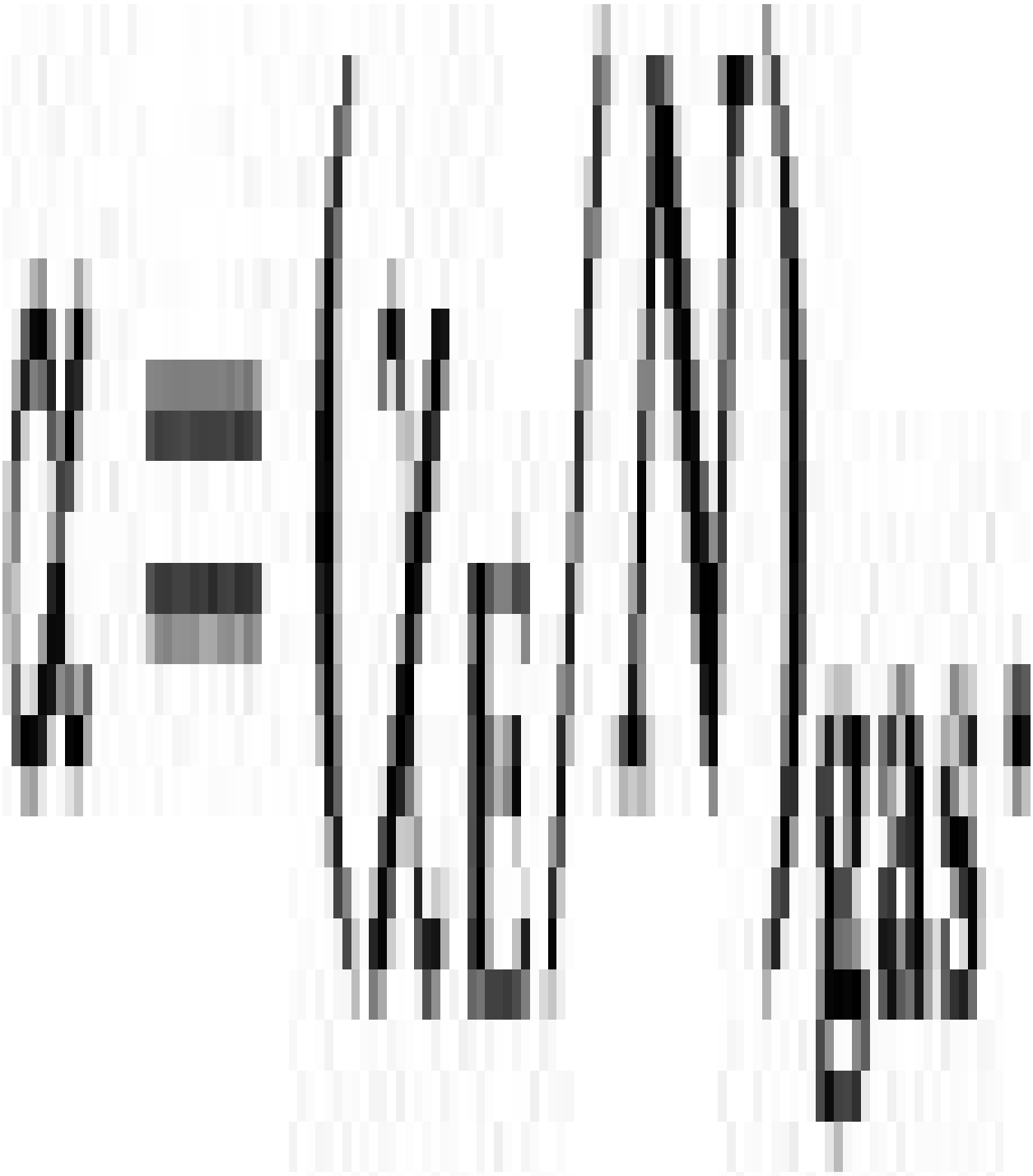
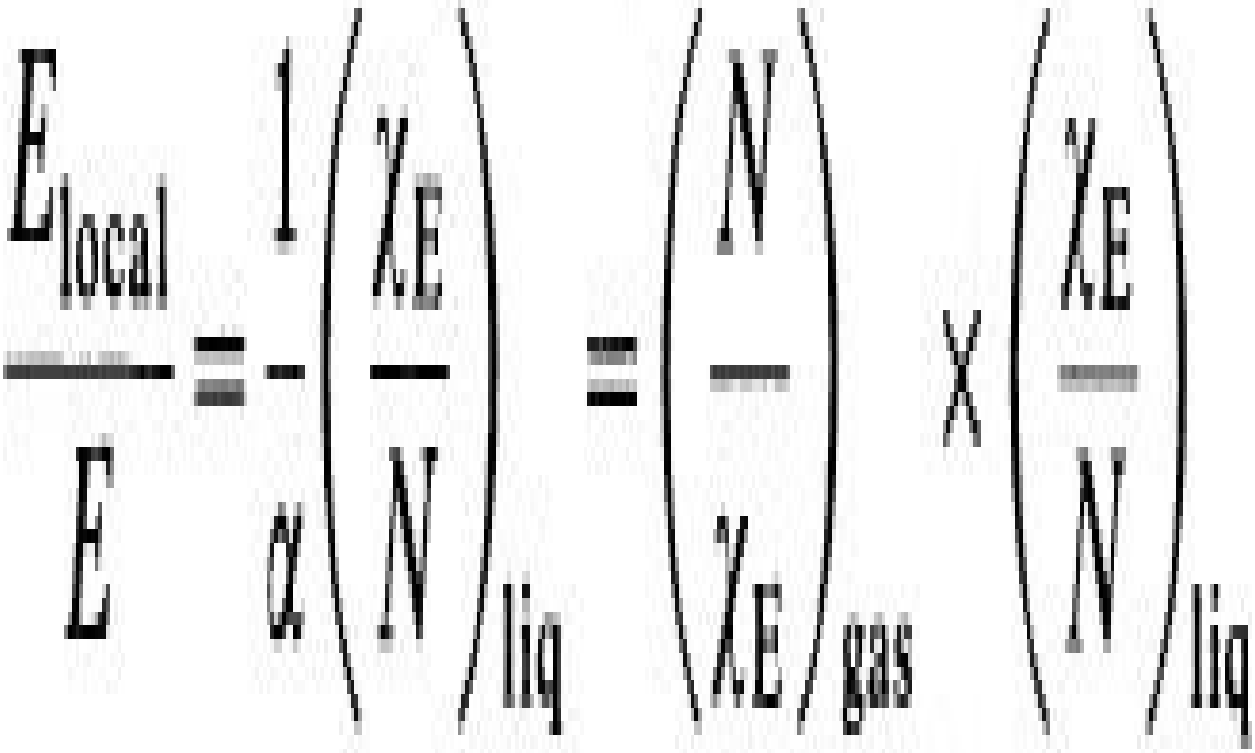


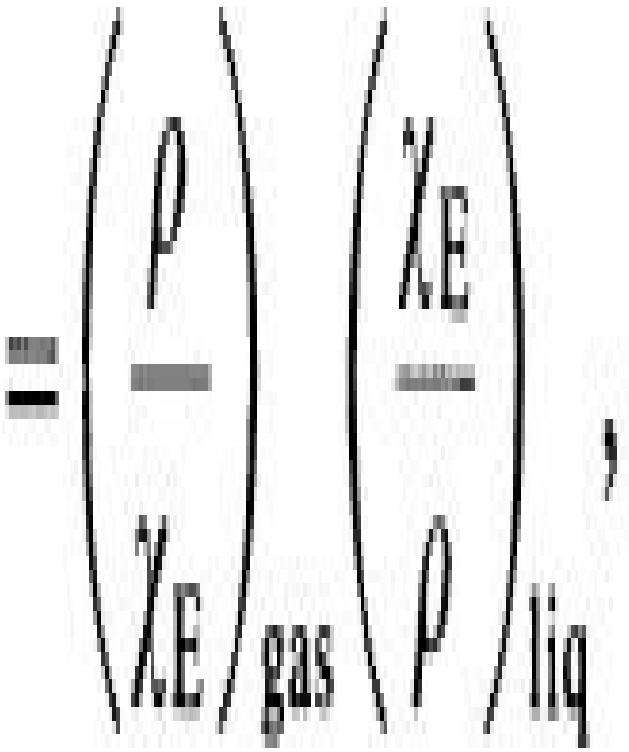
Table 2.3. Electric susceptibilities of gases and liquids.

	Gas			Liquid		$\varepsilon_{real}/\varepsilon$	
Substance	$(x\varepsilon)_{gas}$	$\rho_{gas} \times 10^3$	State	$(x\varepsilon)_{liq}$	ρ_{liq}	$\left(\frac{\rho}{x\varepsilon}\right)_{gas}\left(\frac{x\varepsilon}{\rho}\right)_{liq}$	$1+\frac{3}{2}(x\varepsilon)_{liq}$
He	0.0000684	0.179	2 K	0.0559	0.147	1.00	1.02
O_2	0.000523	1.43	80 K	0.507	1.19	1.165	1.169
CO_2	0.000985	1.96	100 atm	0.610	0.975	1.24	1.20
CS_2	0.0029	3.40		1.64	1.29	1.48	1.55

ρ is the density in g/cm³. The data for gases are at atmospheric pressure.

Comparison of the susceptibility of liquids and gases therefore gives a measure of the local field in the liquid, since





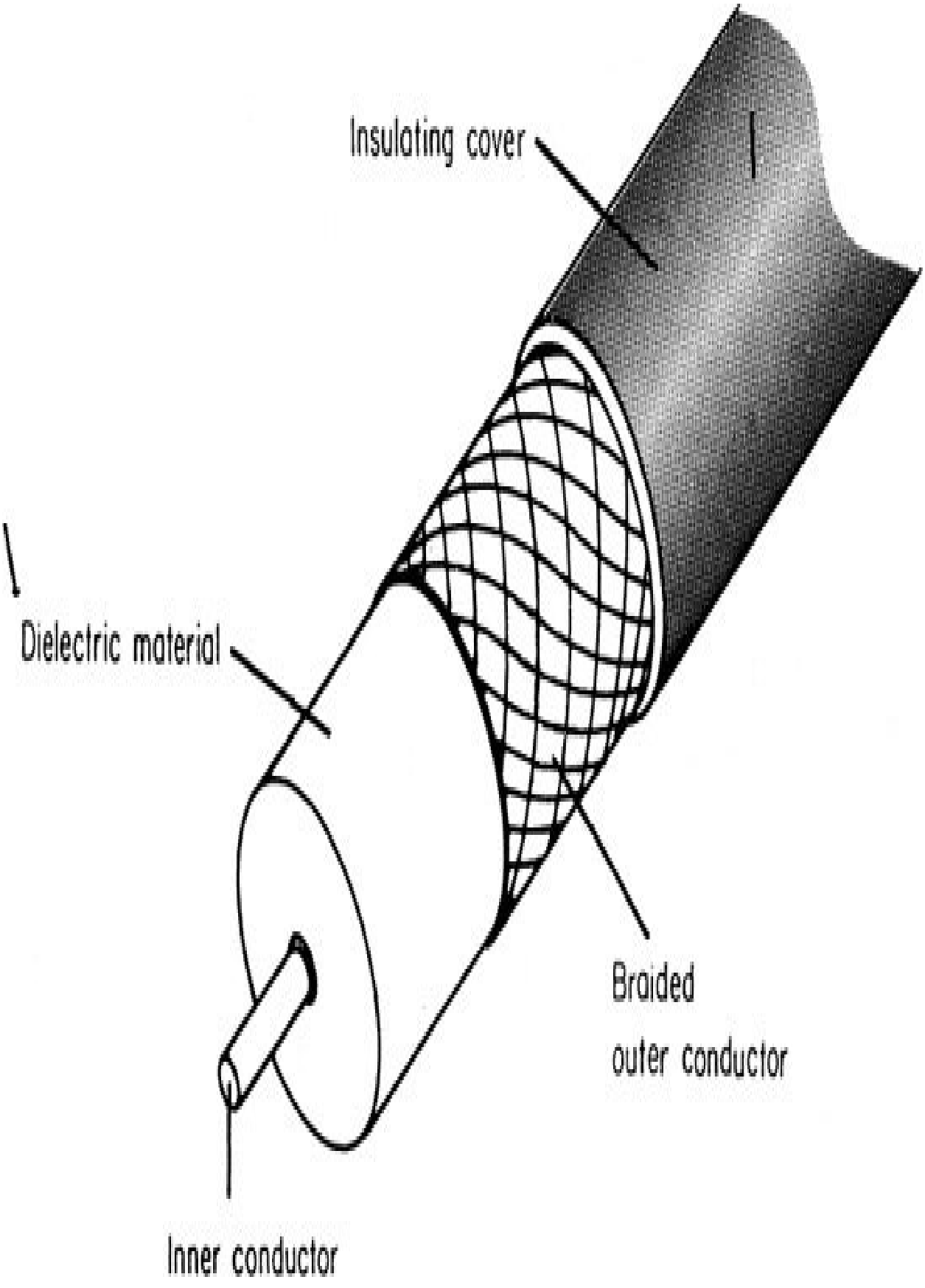
where ρ is the density. Some experimental values are listed in Table 2.3, which compares the value of Elocal/E derived from this expression with the Clausius–Mossotti prediction . The table shows that the Clausius–Mossotti approximation is not bad even when the local field is as much as 50% greater than the macroscopic field.

2.3 MACROSCOPIC FIELDS IN DIELECTRICS

While discussing the atomic basis of dielectric behaviour, it was adequate to restrict attention to uniform polarization. But in practice dielectrics are often found in situations where the polarization is not uniform, either because the dielectric is non-uniform, or because the electric field varies with position even inside a uniform dielectric. For example, the familiar coaxial cable used to connect a television set to its aerial consists of a central conductor separated from an outer cylindrical conductor by a layer of dielectric material, as shown in Figure 2.12. If a potential difference is set up across the conductors, what is the polarization and the electric field in the dielectric?

To answer questions like this, we need to know how induced polarization charges behave in non-uniform fields. In section 2.3.1 it will be shown that in non-uniform fields, as well as induced charges appearing on the surface of a dielectric, there may also be an induced polarization charge distributed throughout the volume of the dielectric. This looks at first sight to be a very awkward complication. Fortunately, by introducing a new vector field called the electric displacement we shall be able to account for the effects of the induced polarization charges automatically, without even working out the polarization charge density. The electric displacement vector is discussed in section 2.3.2, and then in section 2.3.3 we go on to explain how electrostatic fields change at the boundaries between different dielectric media.

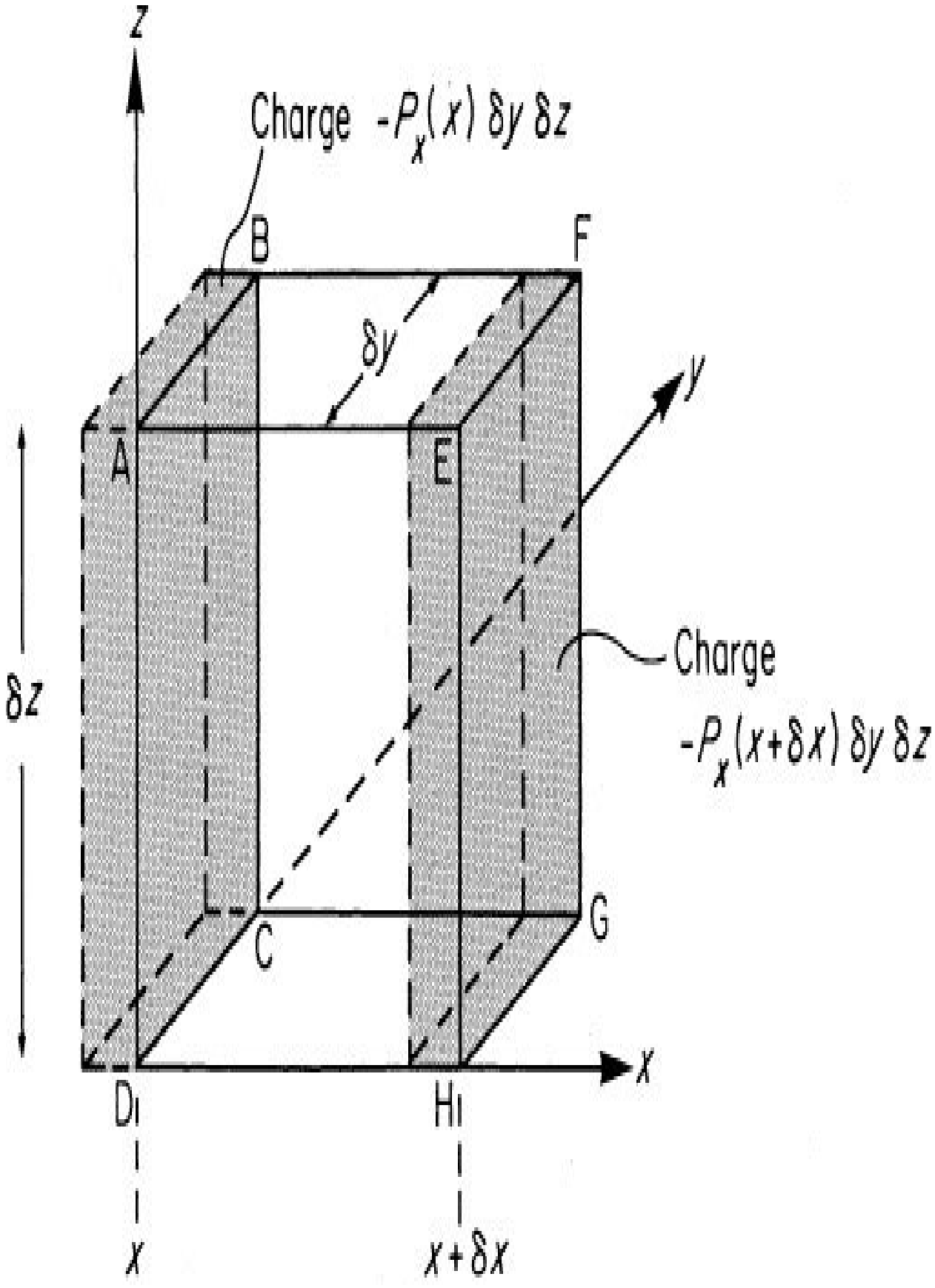
Figure 2.12. A coaxial cable.

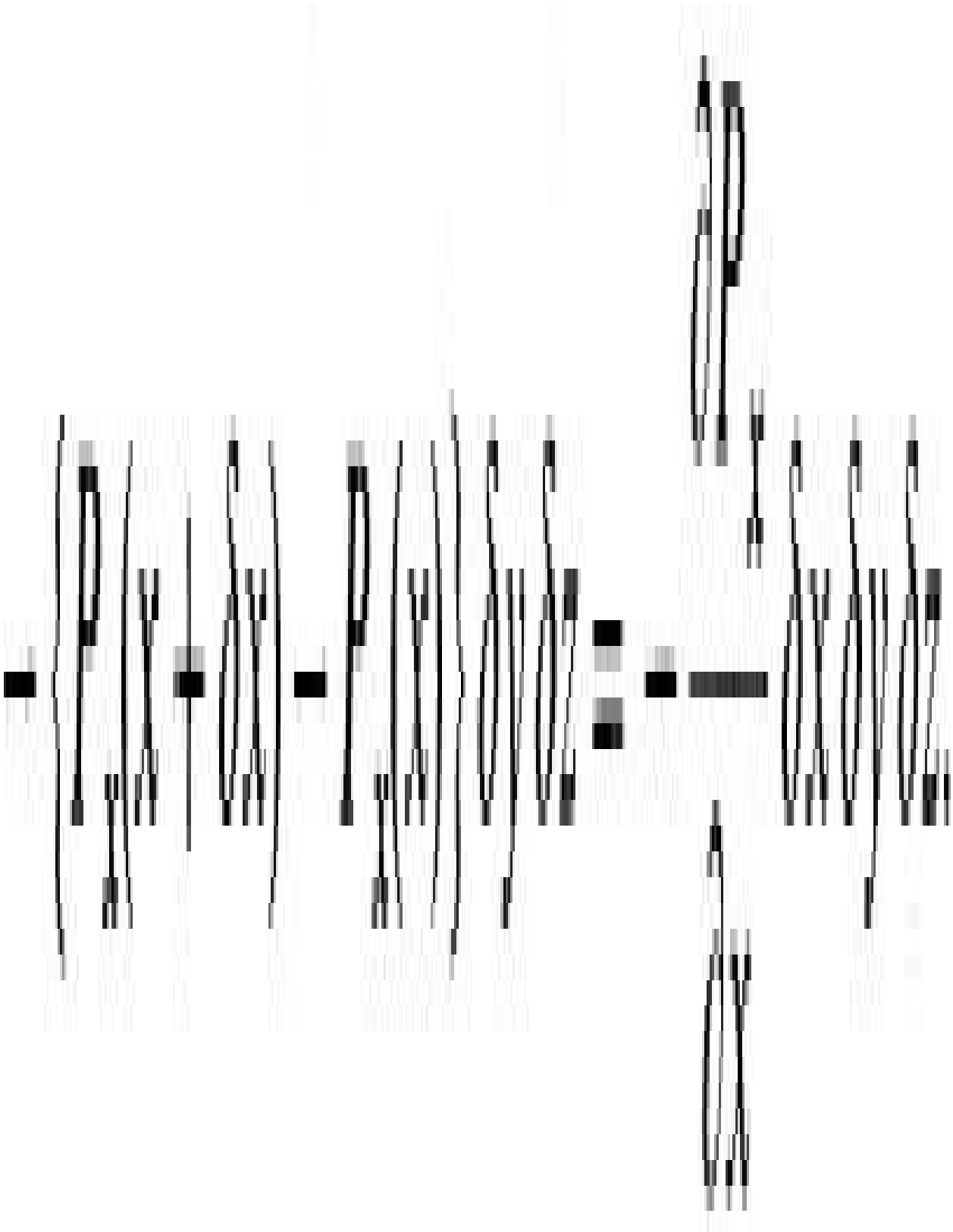

Insulating cover
Dielectric material
Braided
outer conductor
Inner conductor

2.3.1 The volume density of polarization charge

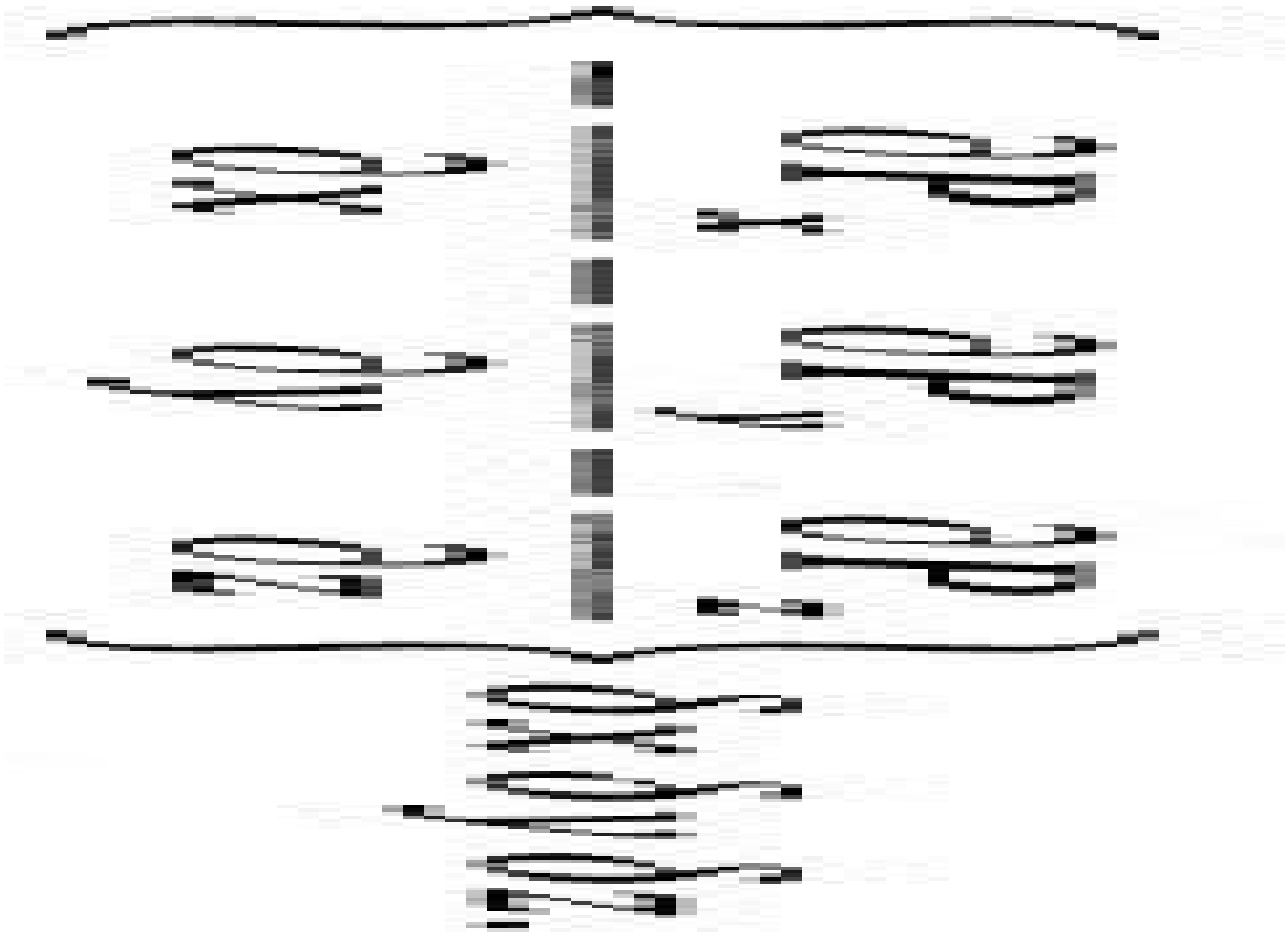
Imagine that the cube of volume $\delta x\, \delta y\, \delta z$ sketched in Figure 2.13 is inside an electrically neutral dielectric. The cube is small on the macroscopic scale, yet still large enough to contain many atoms. Suppose that the cube is placed in an electric field so that it acquires a polarization P. Assume for the moment that the polarization in the cube is in the x-direction, and that it may change with x, but is independent of y and z. The component of polarization $P_x(x)$ is then a function of x only (and is positive), while the components P_y and P_z are zero. When the atoms are polarized, electrons move leftwards out of the box at ABCD, and (using the same argument as in the case of uniform polarization discussed in section 2.1) they carry a charge $-P_x(x)\delta y\, \delta z$. If the polarization has a slightly different value $P_x(x + \delta x)$ at the face EFGH, a charge $-P_x(x + \delta x)\, \delta y\, \delta z$ enters the cube here. The net charge entering the cube through ABCD and EFGH is, to first order in δx,*

Figure 2.13. In a non-uniformly polarized dielectric, different amounts of charge pass through opposite faces of a cube embedded in the dielectric.


z
Charge $-P_x(x)\,\delta y\,\delta z$
B
F
δy
y
A
E
δz
Charge
$-P_x(x+\delta x)\,\delta y\,\delta z$
C
G
x
D
H
x
$x+\delta x$

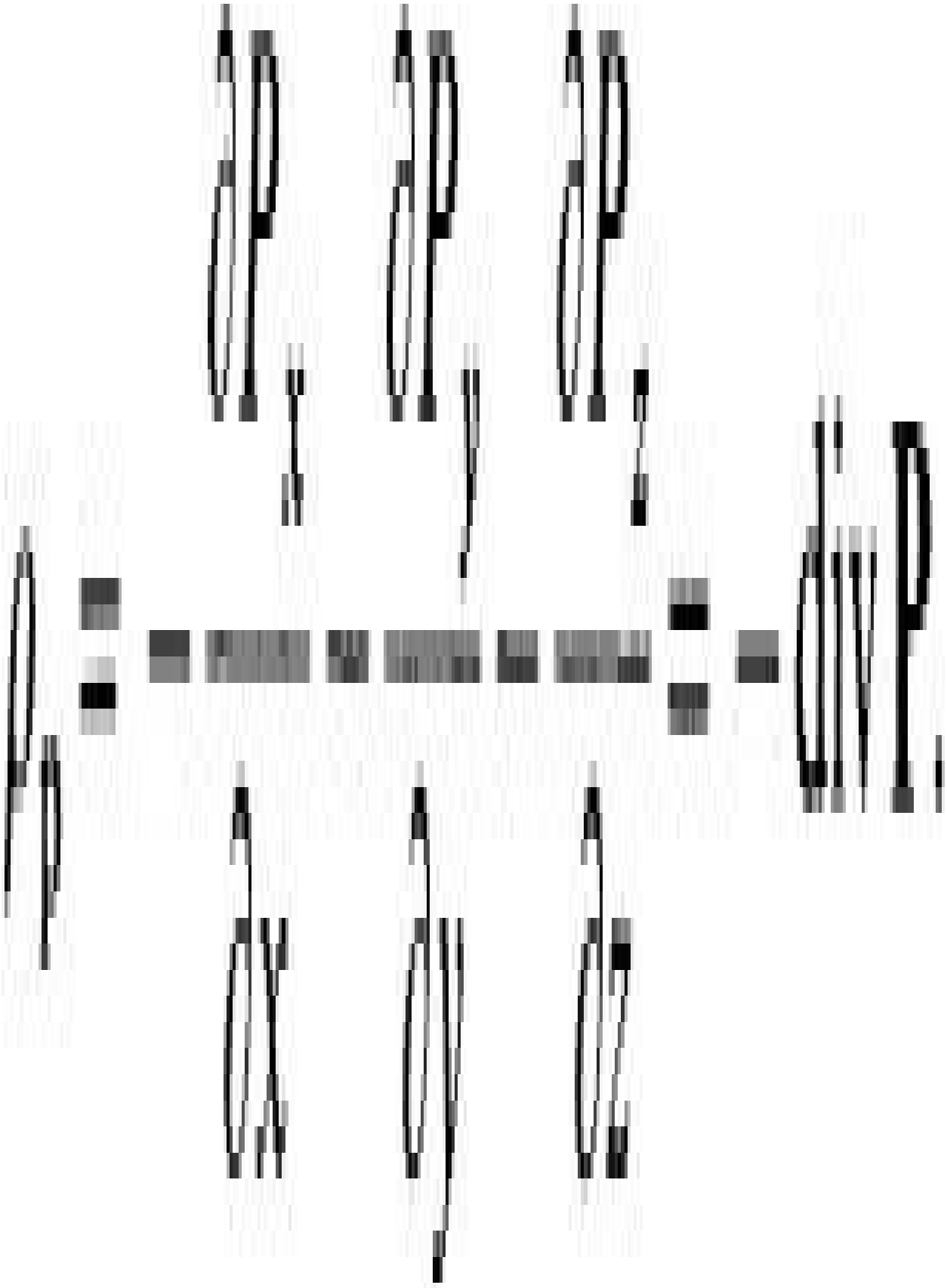


The same expression holds if Px(x) is negative, since the electrons then move in the opposite direction. In general there may be similar contributions from components of P in the y and z directions, and the total polarization charge acquired by the cube is



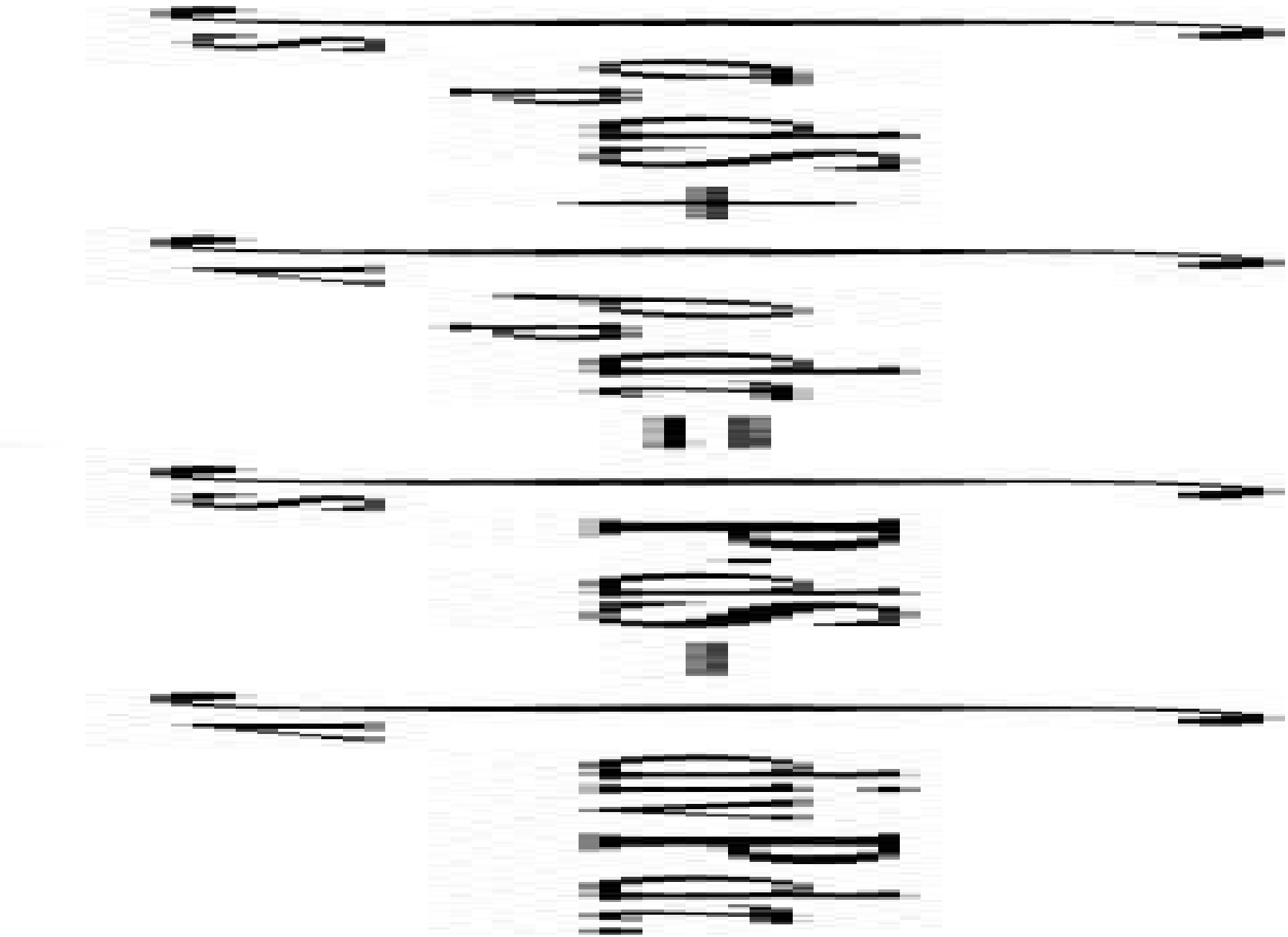
In other words there is a macroscopic polarization charge density

(2.17)



Equation (2.17) is similar to Gauss' law which relates the electric field to the total charge density. Note however the minus sign, which arises because of the convention that dipole moments are drawn in the direction from negative to positive charge. Field lines of the vector P terminate on polarization charges, starting from negative charges and ending on positive charges.

The appearance of the polarization charge density is consistent with the requirement that the total polarization charge carried by a piece of dielectric should be zero. The surface charge on an element dS on the surface of the dielectric is P·dS, and, summing surface and volume charges, the total polarization charge is



The divergence theorem (Equation (1.21)) tells us that the right-hand side is identically zero.

2.3.2 The electric displacement vector

In the general case, dielectric material may not be electrically neutral even when unpolarized. If the dielectric carries a charge density ρf of ‘free’ charges, representing a net surplus or deficit of electrons in the atoms of the dielectric then the total charge density is

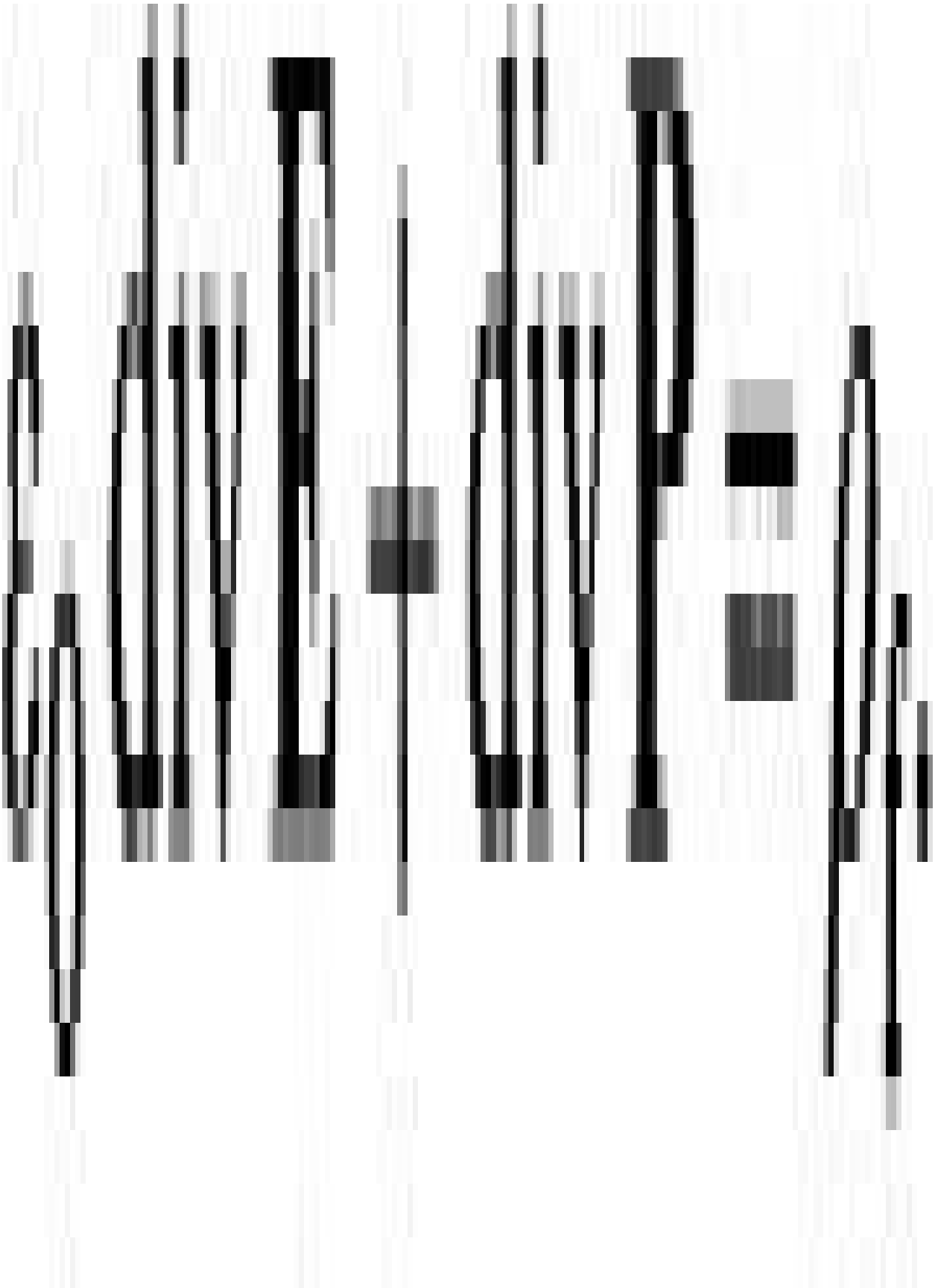
(2.18)

It must be emphasized that both terms in this equation represent real physical charge. Nevertheless, we shall find the distinction between free charge and polarization charge to be a useful one, particularly when we come to consider the energy associated with fields in dielectric material. The macroscopic electric field E is related to the total charge density, and in matter Gauss’ law becomes

$$\operatorname{div}\mathbf{E} = \rho/\varepsilon_0 = \frac{1}{\varepsilon_0}(\rho_f + \rho_p)$$

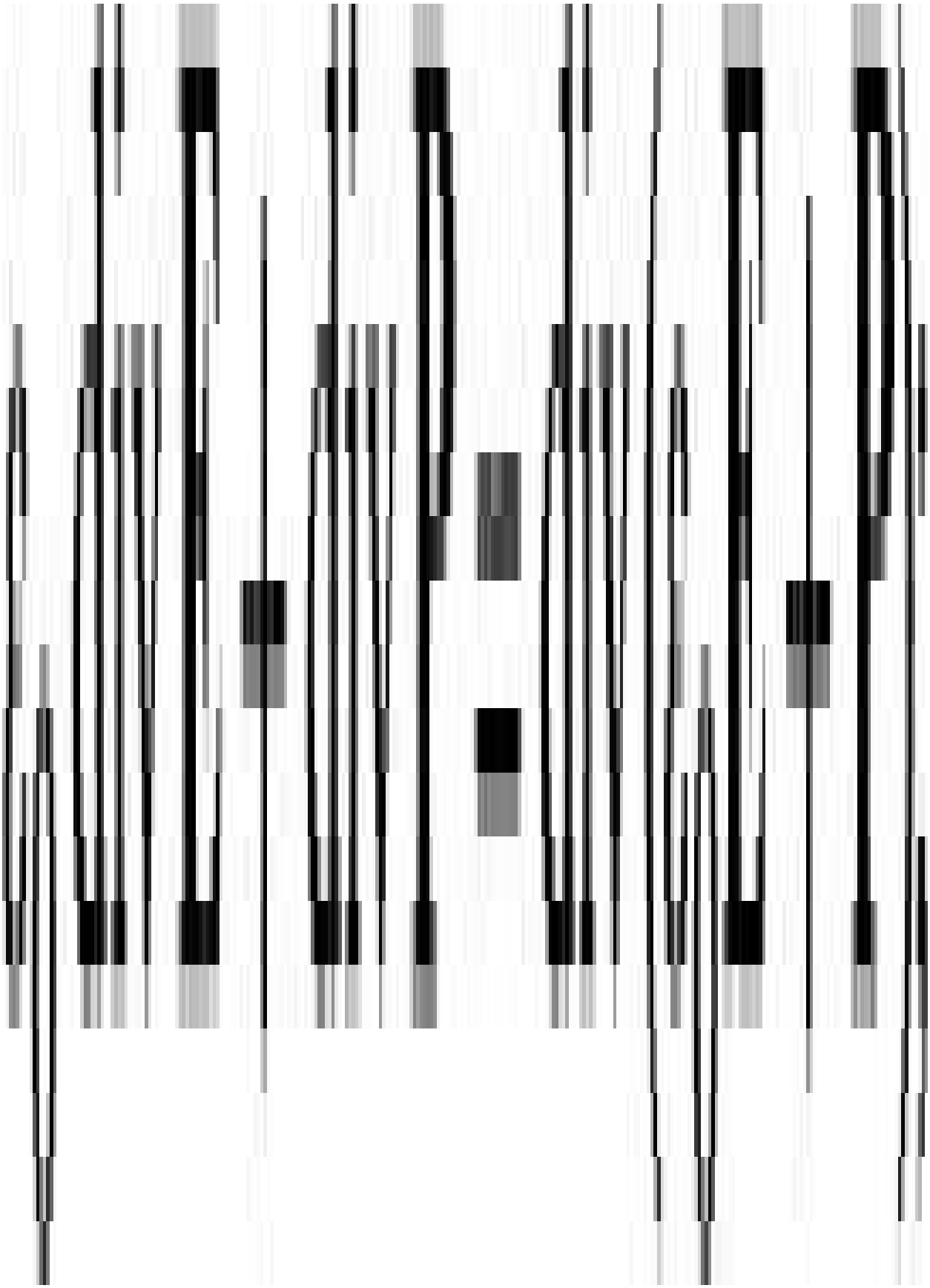
$$= \frac{1}{\varepsilon_0}(\rho_f - \operatorname{div}\mathbf{P}),$$

Rearranging this equation we can write

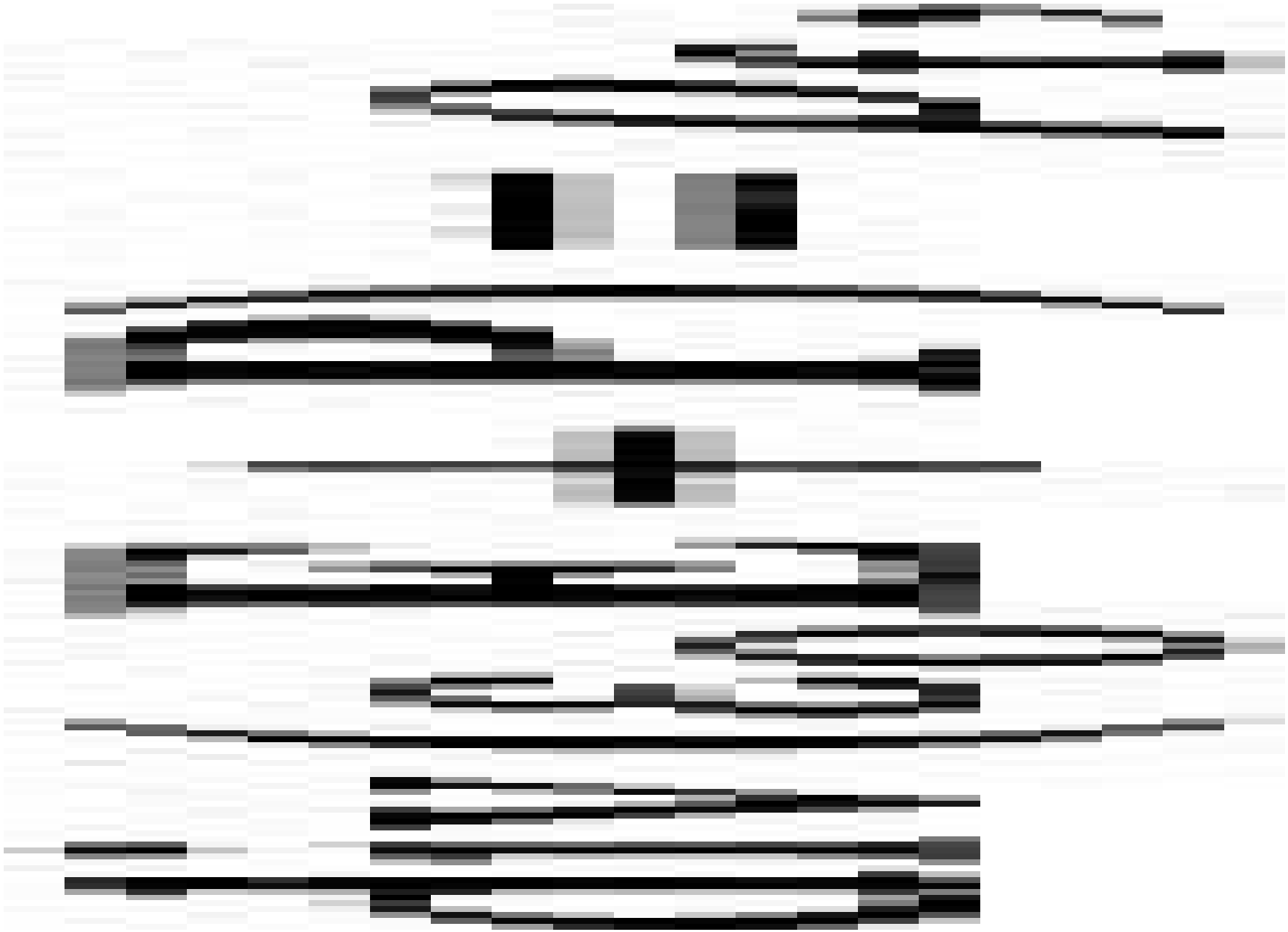


Note that since P is proportional to E, div P ≠ 0 only if $\rho_f \neq 0$. Because $\rho_p = -$ div P, we find that a polarization charge density ρ_p is found only in regions where there is also a free charge density.

It can easily be verified, by writing out the equation in full in terms of the components of the vectors, that



hence



or

(2.19)

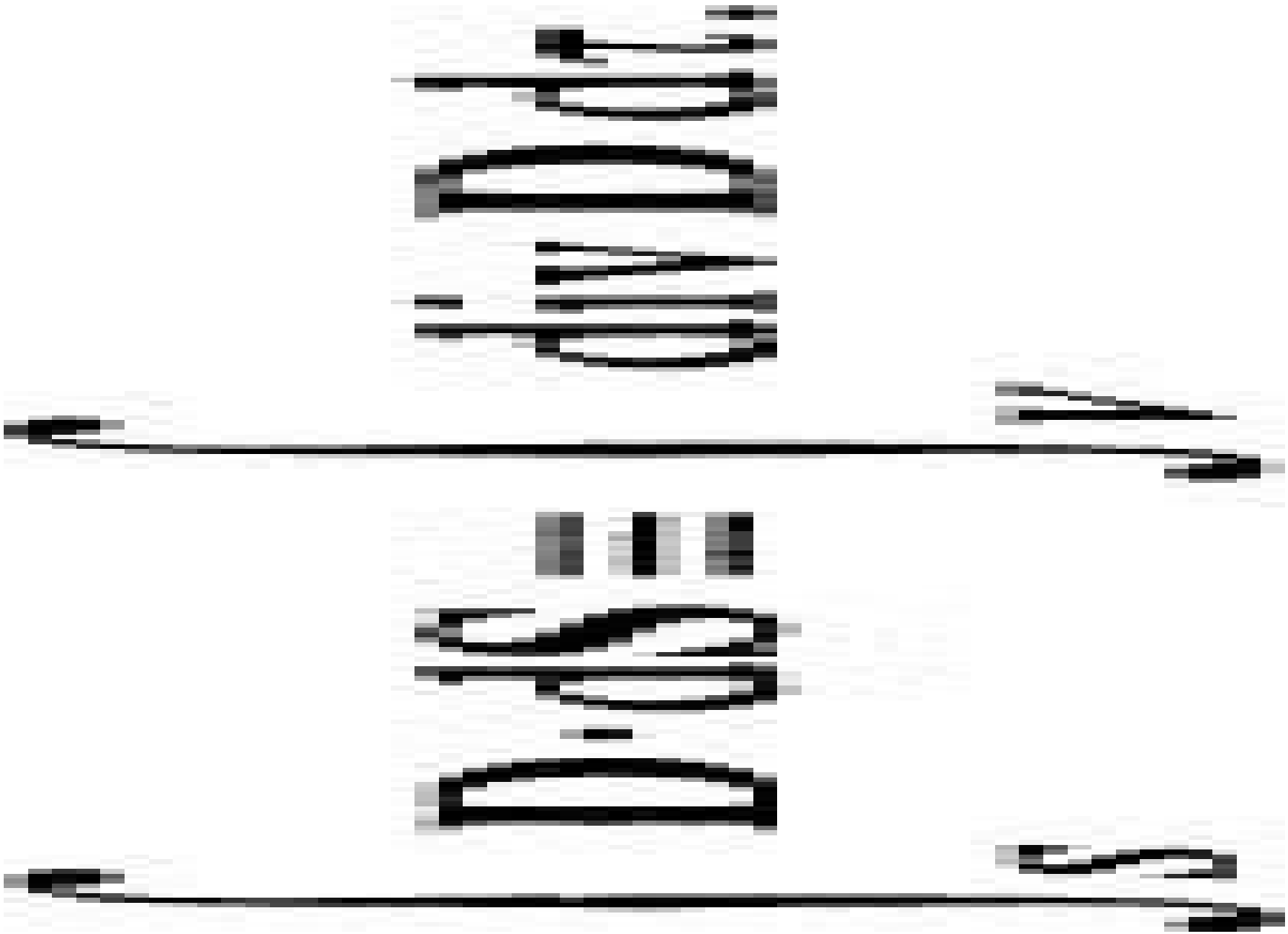
Here we have introduced a new vector field D(r), which is called the electric displacement, and is defined as

(2.20)

The vector field D has the same dimensions as P, namely dipole moment per unit volume. Since P = χEε0E the electric displacement can also be written as

(2.21)

Equation (2.19) is still really Gauss' law, now modified in such a way that the effects of polarization charge are automatically included. (Gauss' law looks a bit different in that it contains a factor l/ε0 on the right-hand side. This is simply because we have chosen to give D and E different dimensions.) Gauss' law can be expressed alternatively in the integral form ∫sE·dS = (1/ε0)∫vρ dτ. Similarly there is an integral relation for D which is equivalent to Equation (2.19). The divergence theorem tells us that



Substituting for div D from Equation (2.19)

(2.22)

or in words,

Flux of D out of a closed surface S = Total free charge enclosed within S.

Unlike the electric field E (which is the force acting on unit charge) or the polarization P (the dipole moment per unit volume), the electric displacement D has no clear physical meaning. The only reason for introducing it is that it enables one to calculate fields in the presence of dielectrics without first having to know the distribution of polarization charges.

Let us illustrate this by applying Equation (2.22) to the example of the coaxial cable shown in section in Figure 2.14. The annular region between the inner and outer conductors is filled with a uniform dielectric material of relative permittivity ε. The two conductors make up a capacitor, and when a potential difference is maintained between them, induced charges appear on the conductor surfaces. Suppose that there is a surface charge density σf on the inner conductor, which has a radius a, so that the total charge on a one metre length is $Q = 2\pi a\sigma_f$. The electric displacement is obviously cylindrically symmetrical, and its magnitude depends only on the distance r from the axis. The vector D points radially outwards and the flux of D out of a cylindrical surface S of radius r and length one metre is

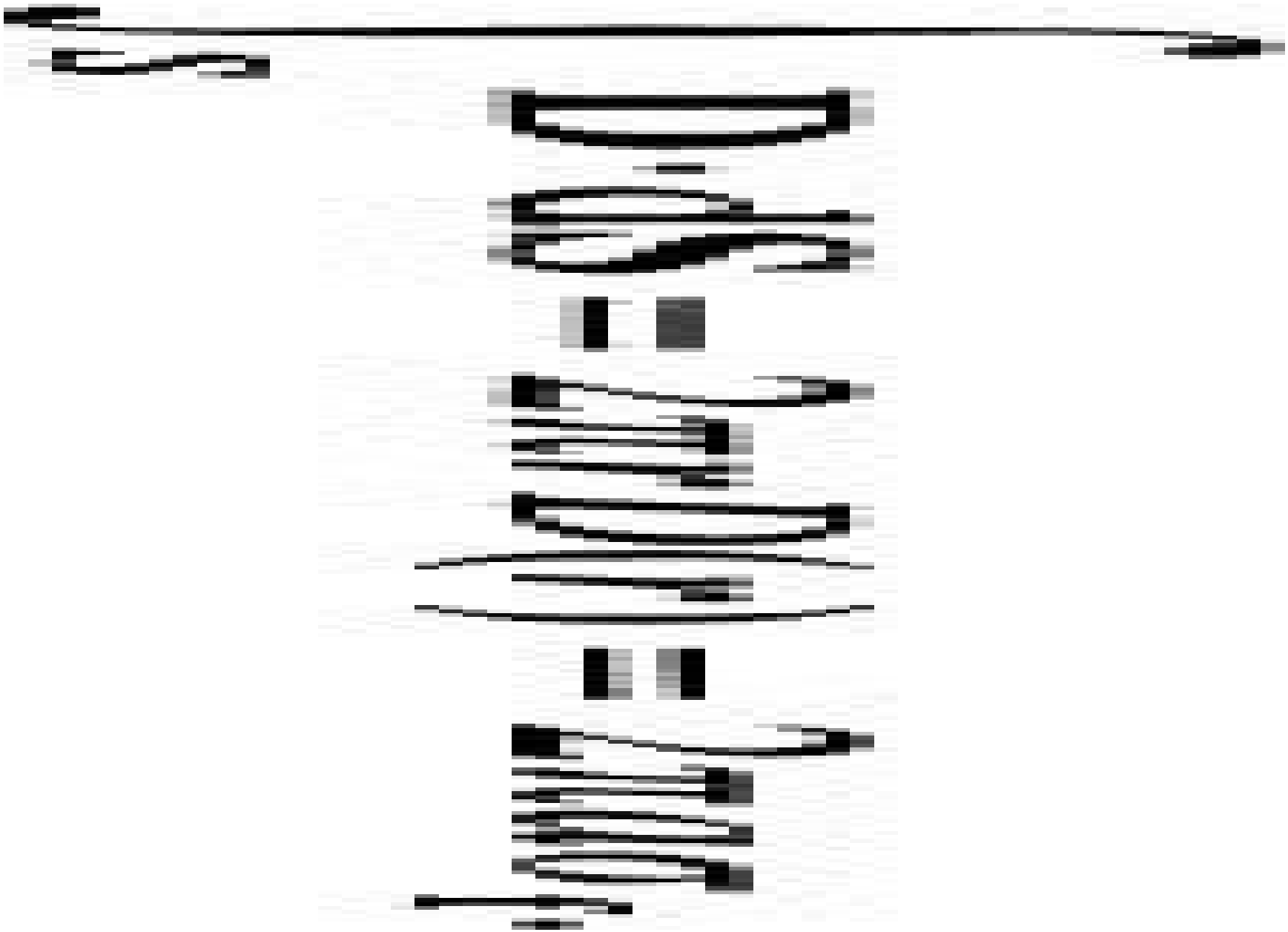
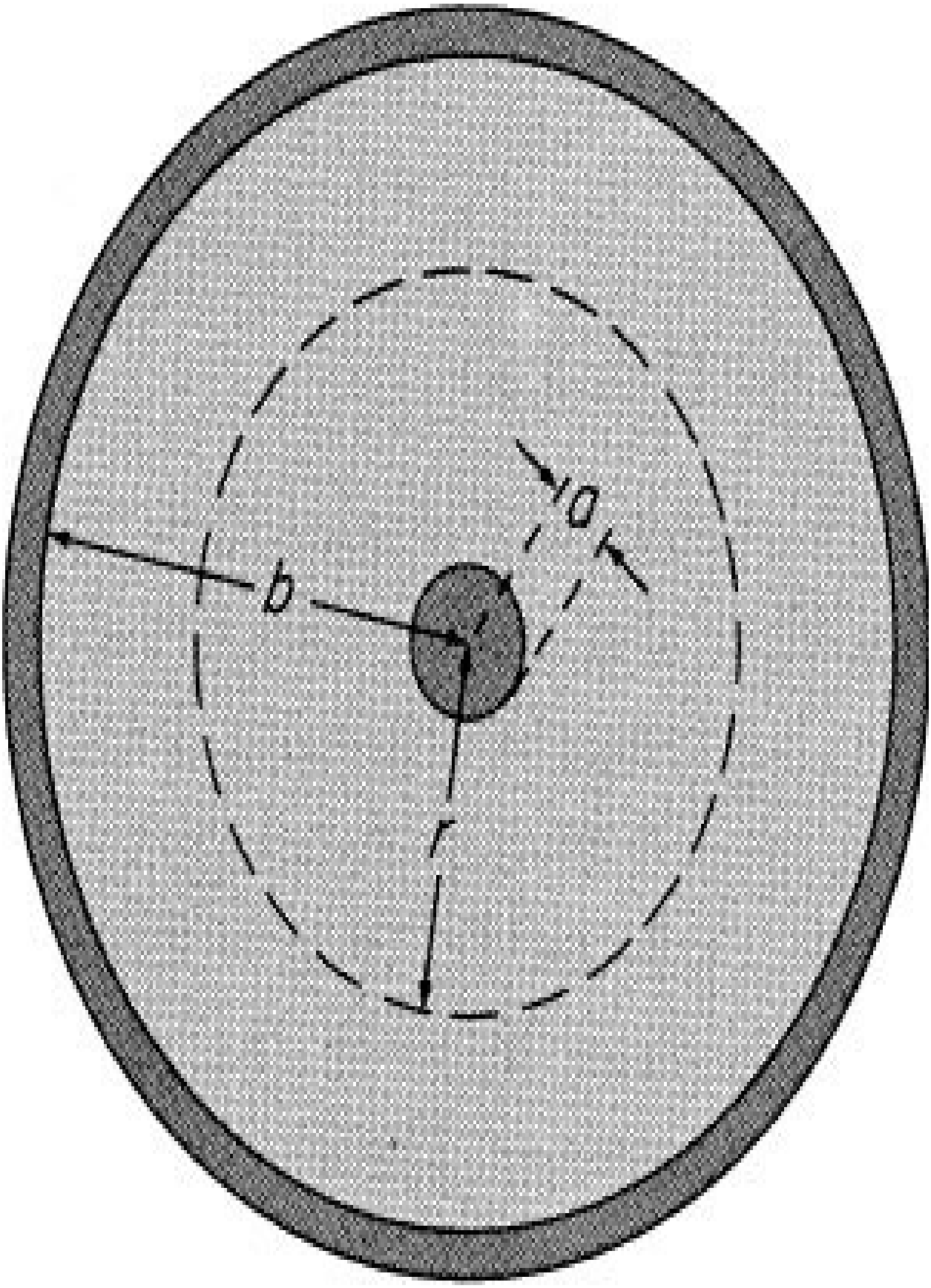
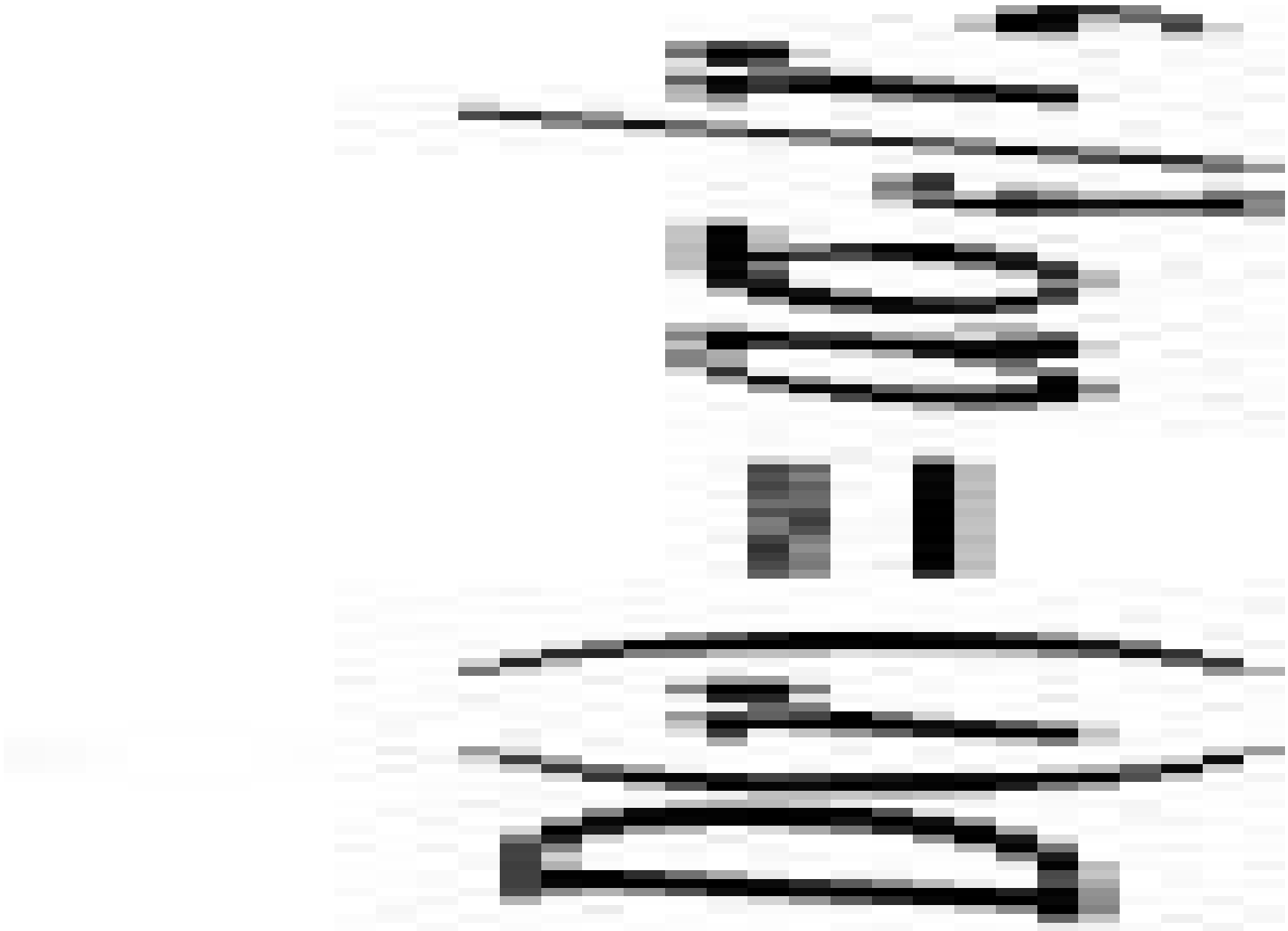


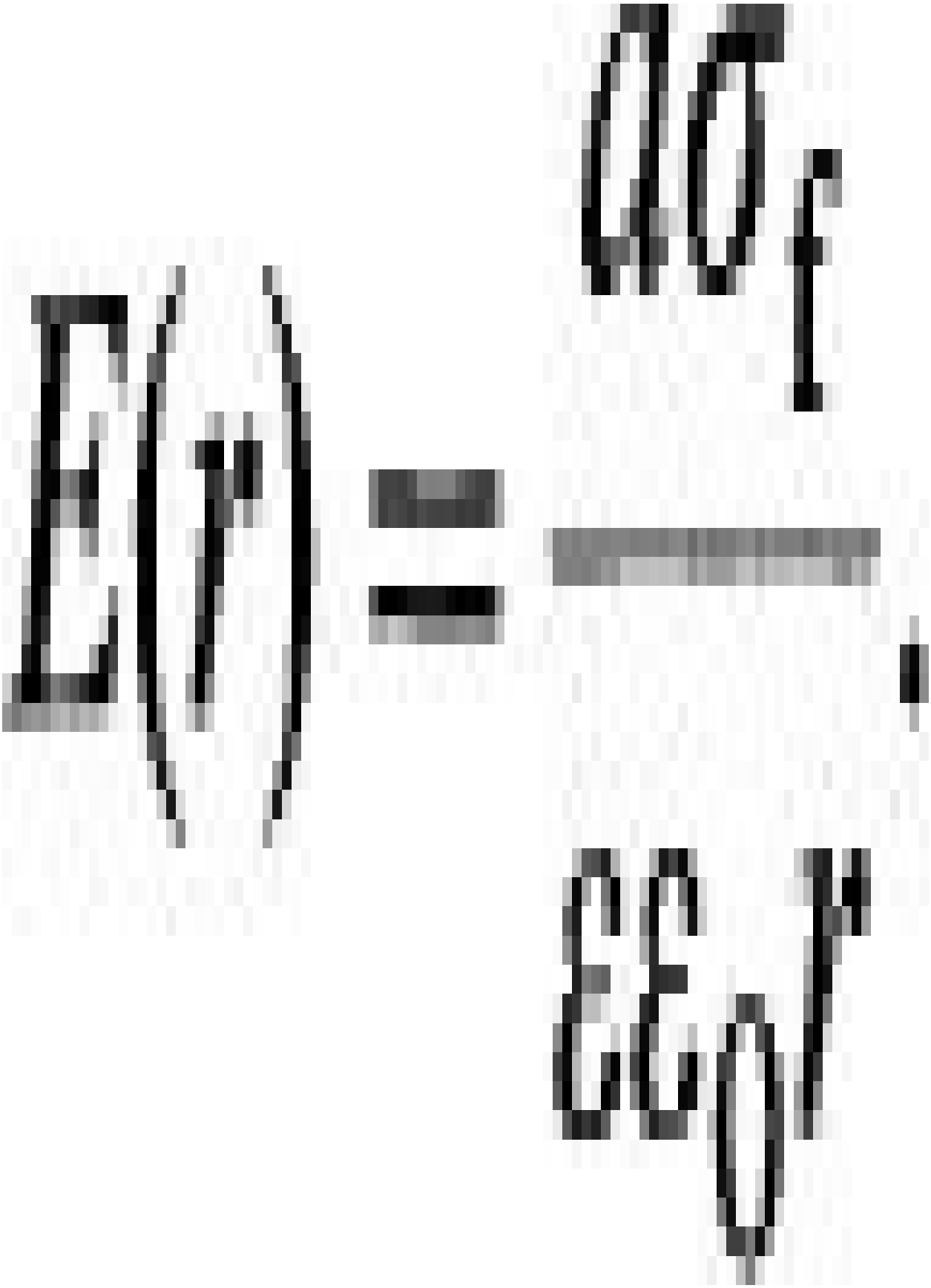
Figure 2.14. Cross-section of a coaxial cable.


b

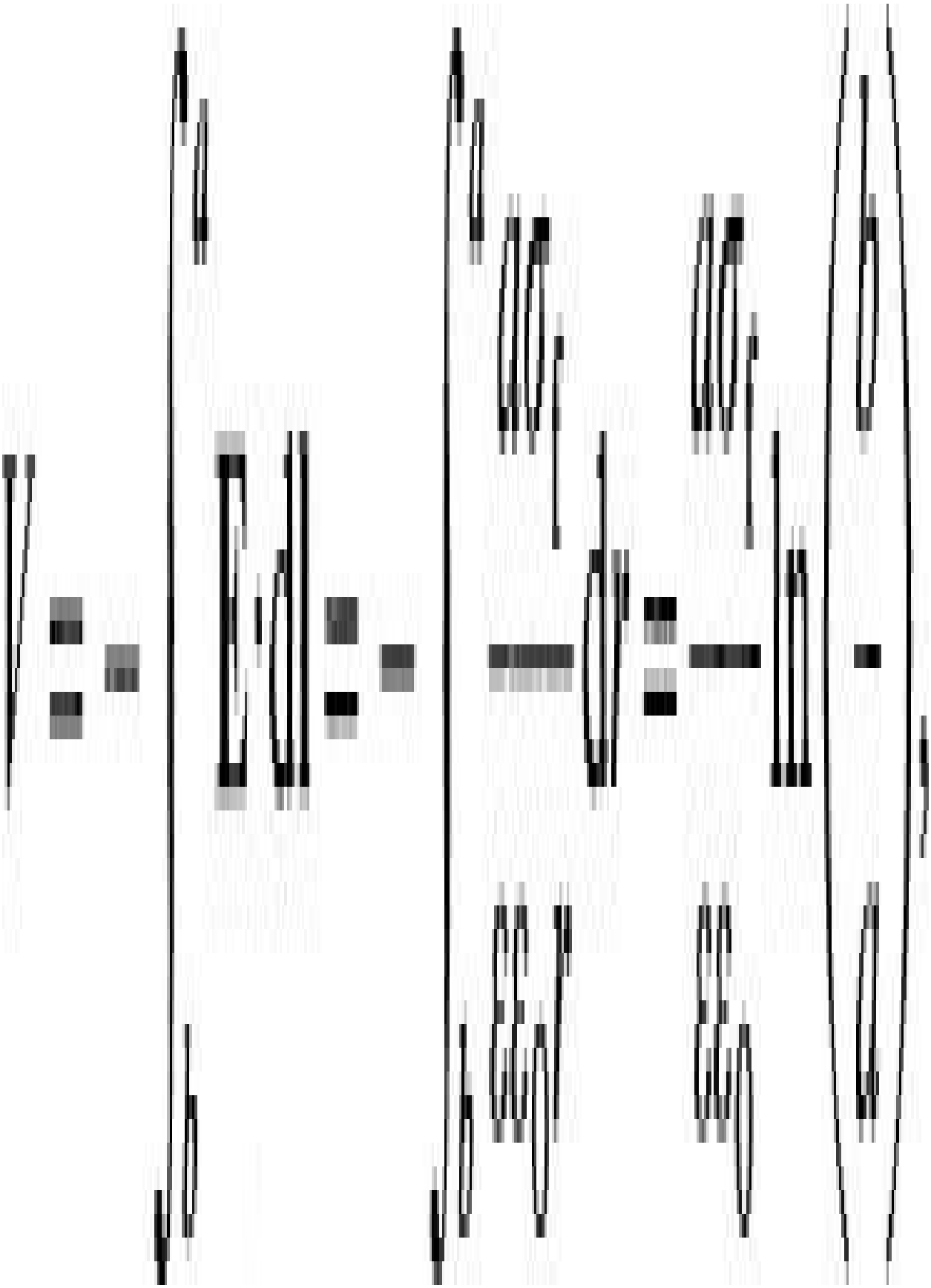
Hence



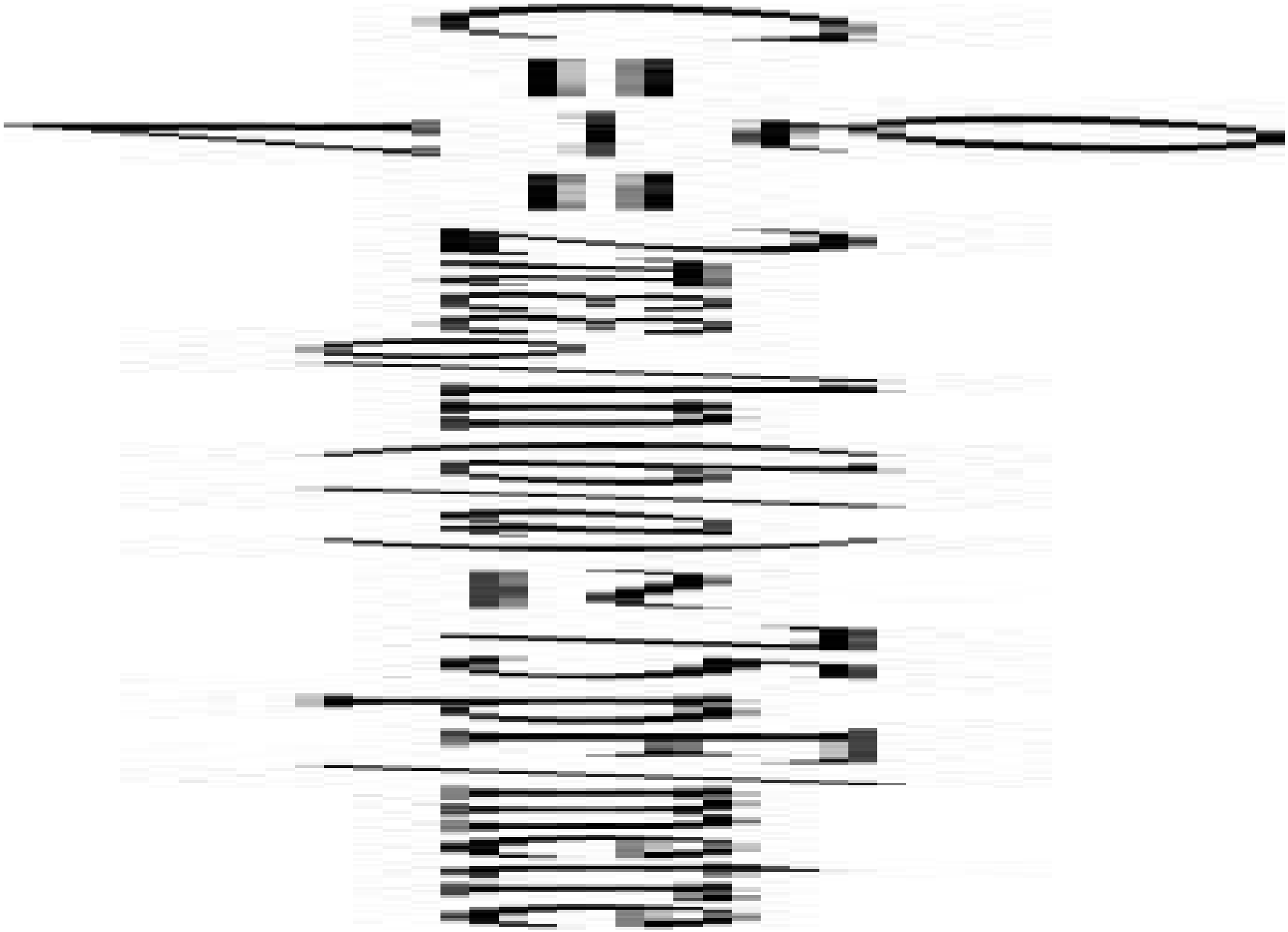
and from equation (2.21)



The potential difference between the inner and outer conductor is



and the capacitance per unit length of the cable is



for a typical coaxial cable with polythene as the dielectric ($\varepsilon = 23$) and a = 0.5 mm, b = 2.5 mm. The capacitance per unit length is an important property of a coaxial cable, which one needs to know when discussing its performance in the transmission of high-frequency signals.

The electric field in the cable varies as 1/r, just as it would do if there were no dielectric between the conductors. Whenever the relative permittivity is the same throughout the whole of the region where the field is non-zero, Equation (2.19) simplifies to

div D = ε₀ div E = ρ_f

or div E = ρ_f/ε₀.

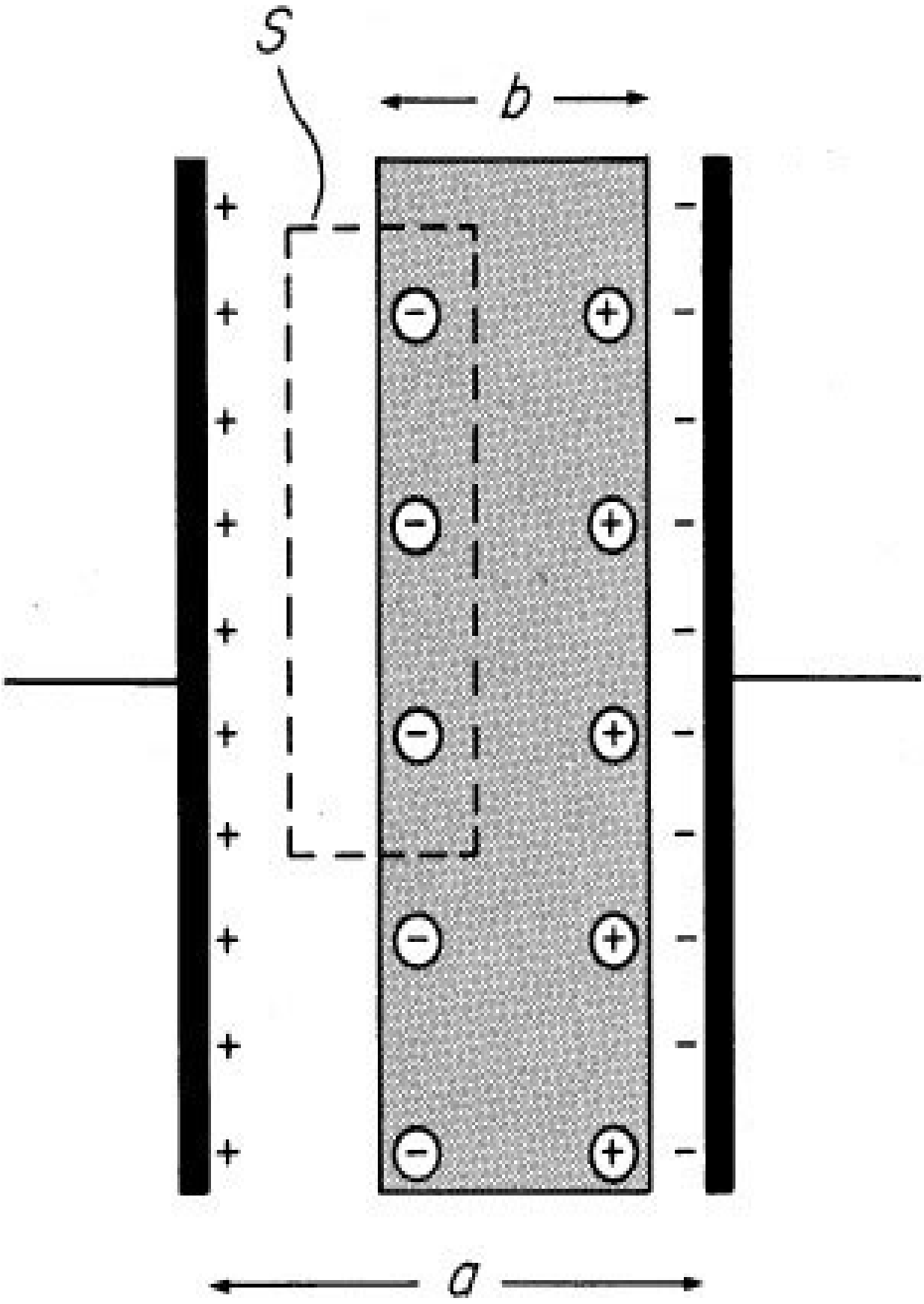
Apart from the factor ε, this is the equation satisfied by the same distribution of free charge ρf in vacuum (Equation (1.20)). The presence of the dielectric does not alter the shape of the electric field, but simply reduces its magnitude by the factor ε.

2.3.3 Boundary conditions for D and E

When the space near a set of charges contains dielectric, but is not completely filled by a single uniform dielectric material, then the electric field no longer has the same form as in vacuum. Suppose, for example, that a slab of dielectric, which carries no free charges, is placed in a parallel plate capacitor, but that the thickness b of the slab is less than the distance a between the capacitor plates. When there is a potential difference between the plates, polarization charges appear at the free surface of the dielectric. There is therefore a discontinuity in the electric field: by Gauss' law, the fluxes of the field E through opposite faces of the box S shown in Figure 2.15 are not equal, since the box encloses polarization charges. But the flux law for the electric displacement D (Equation (2.22)) involved only free charge. There is no free charge inside the box, and the fluxes of D across its opposite faces are equal. In other words, the vector D obeys the simple boundary condition that D is continuous across a surface perpendicular to D, if there is no free charge on the surface.

It is now easy to work out the electric field distribution in the capacitor and hence derive its capacitance. The electric displacement D is uniform throughout the capacitor. In the region where there is no dielectric material, the electric field is $E1 = D/\varepsilon 0$. The dielectric has a uniform polarization P, and inside it the electric field is $E2 = D/\varepsilon\varepsilon 0$. The potential difference between the plates is thus

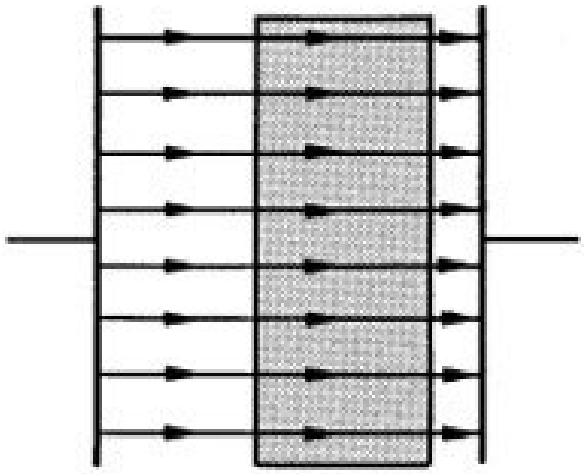
Figure 2.15. There is no change in D across the dielectric surface enclosed by the box S.


S
b
a

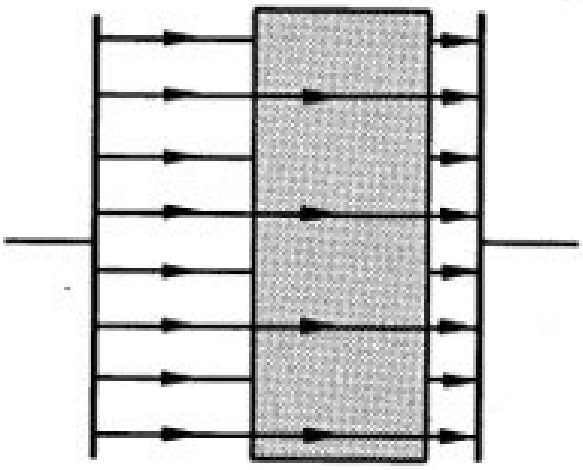
The charge densities on the conducting plates are ± ε0E1, and if their area is A, the plates carry total charges ± Q = ± ε0E1A. The capacitance is therefore C = Q/V = ε0A/(a – b + b/ε). The fields D, E and P in the capacitor are illustrated in Figure 2.16.

The parallel plate capacitor is a specially simple case, but we can apply similar arguments to find the boundary conditions at any dielectric boundary. Imagine a disc enclosing part of the boundary surface between two media of relative permittivity ε1 and ε2, as indicated in Figure 2.17. The thickness of the disc is allowed to tend to zero, so that the only contributions to the outward flux of D from the disc come from its flat faces. These faces have areas represented by the outward normals δS1 and δS2. If there is no free charge inside the disc, the total outward flux of D is zero:

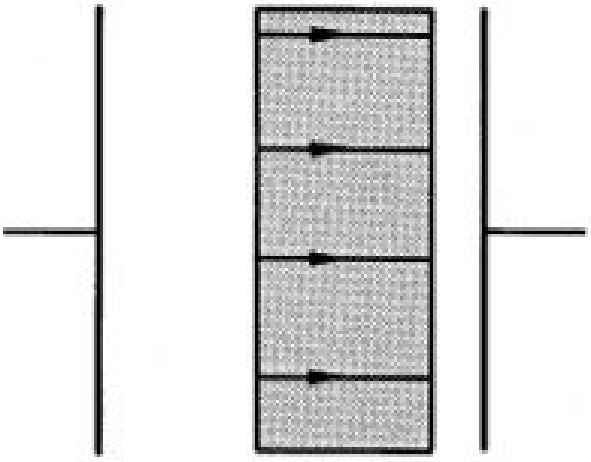
Figure 2.16. Lines of D, E and P in a parallel plate capacitor which is partially filled with dielectric material.



D is constant throughout the capacitor.

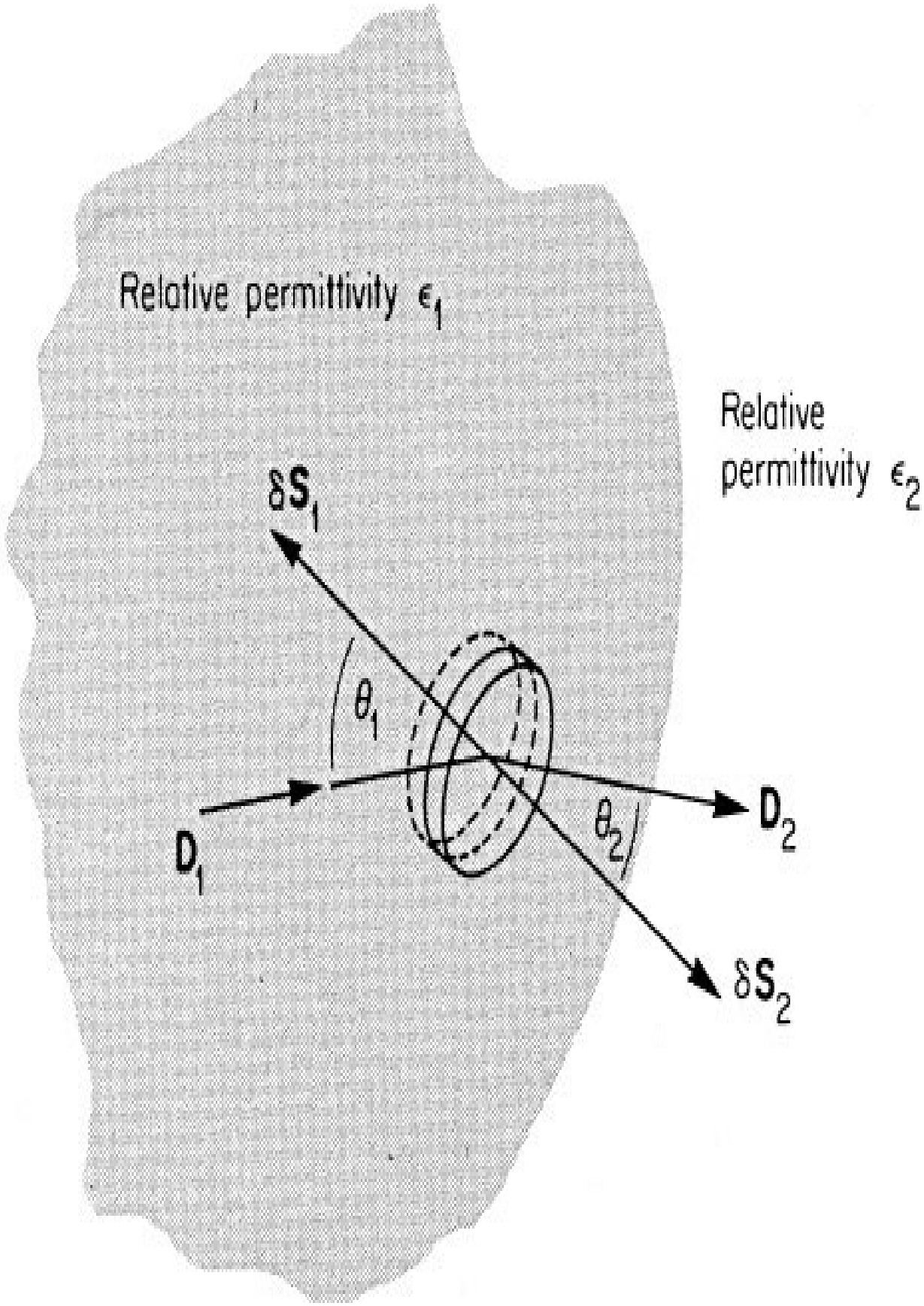


E is *reduced* inside the dielectric, where there are fewer lines.



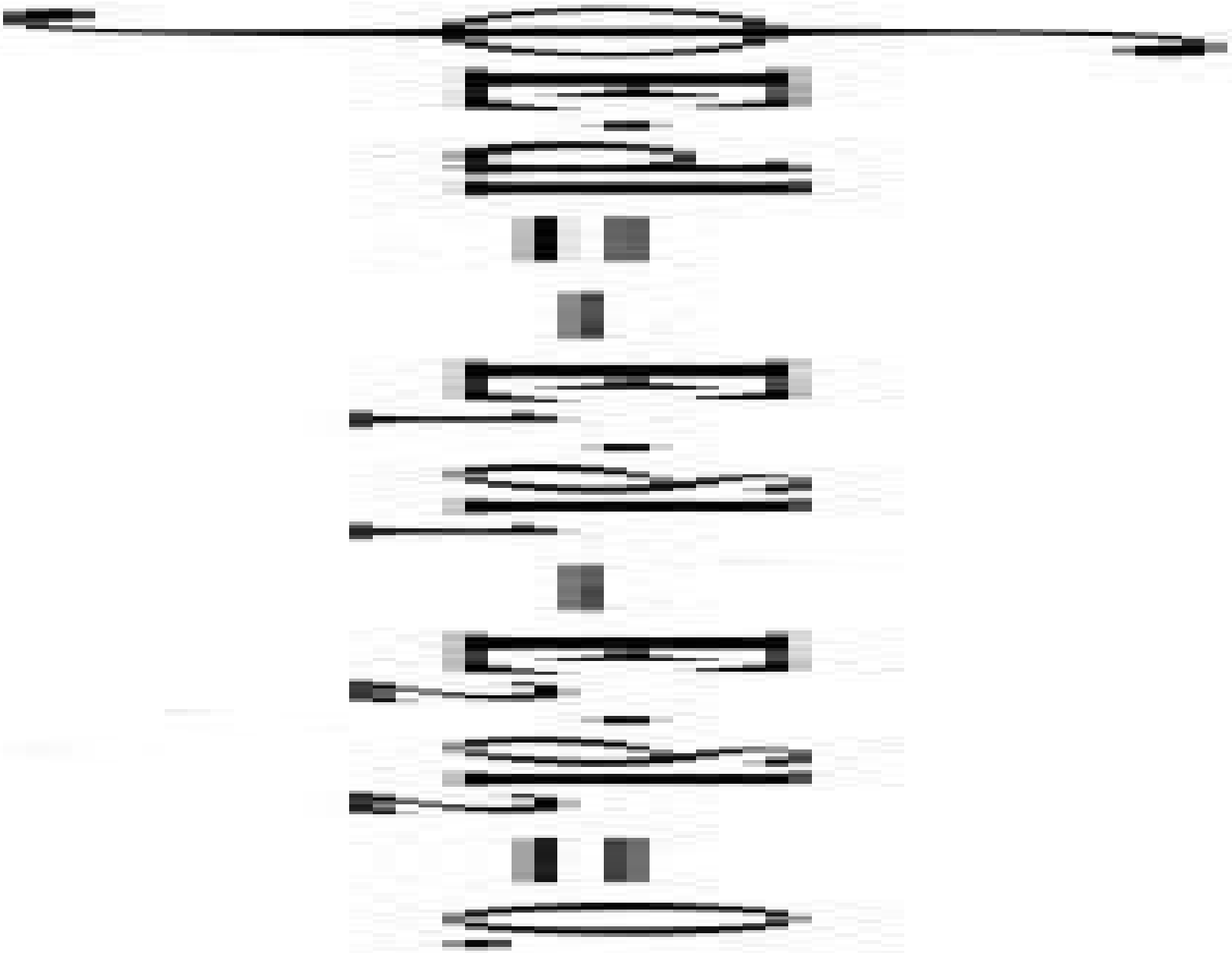
P is zero except inside the dielectric.

Figure 2.17. The change of D at a dielectric boundary.


Relative permittivity ϵ_1
Relative permittivity ϵ_2
δS_1
θ_1
D_1
D_2
θ_2
δS_2

Writing $D1_{\perp}$ and $D2_{\perp}$ for the components of D1 and D2 perpendicular to the boundary, and bearing in mind that $\delta S1$ and $\delta S2$ are pointing in opposite directions, it follows that $D1_{\perp} = D2_{\perp}$, i.e. $D_{\perp}$ is continuous across the boundary. (At a boundary carrying a surface charge density σf there is a discontinuity in $D_{\perp}$ of magnitude σf.)

A different boundary condition applies to the components of the fields parallel to the boundary; in this case it is the component of the electric field which is continuous. Consider the rectangular loop shown in Figure 2.18 with sides $\delta I1$ and $\delta I2$ lying close to the boundary between media with relative permittivities $\varepsilon 1$ and $\varepsilon 2$. Since the electrostatic field is conservative, no work is done in taking a test charge around the loop. Making the short sides vanishingly small, the work done on unit charge is



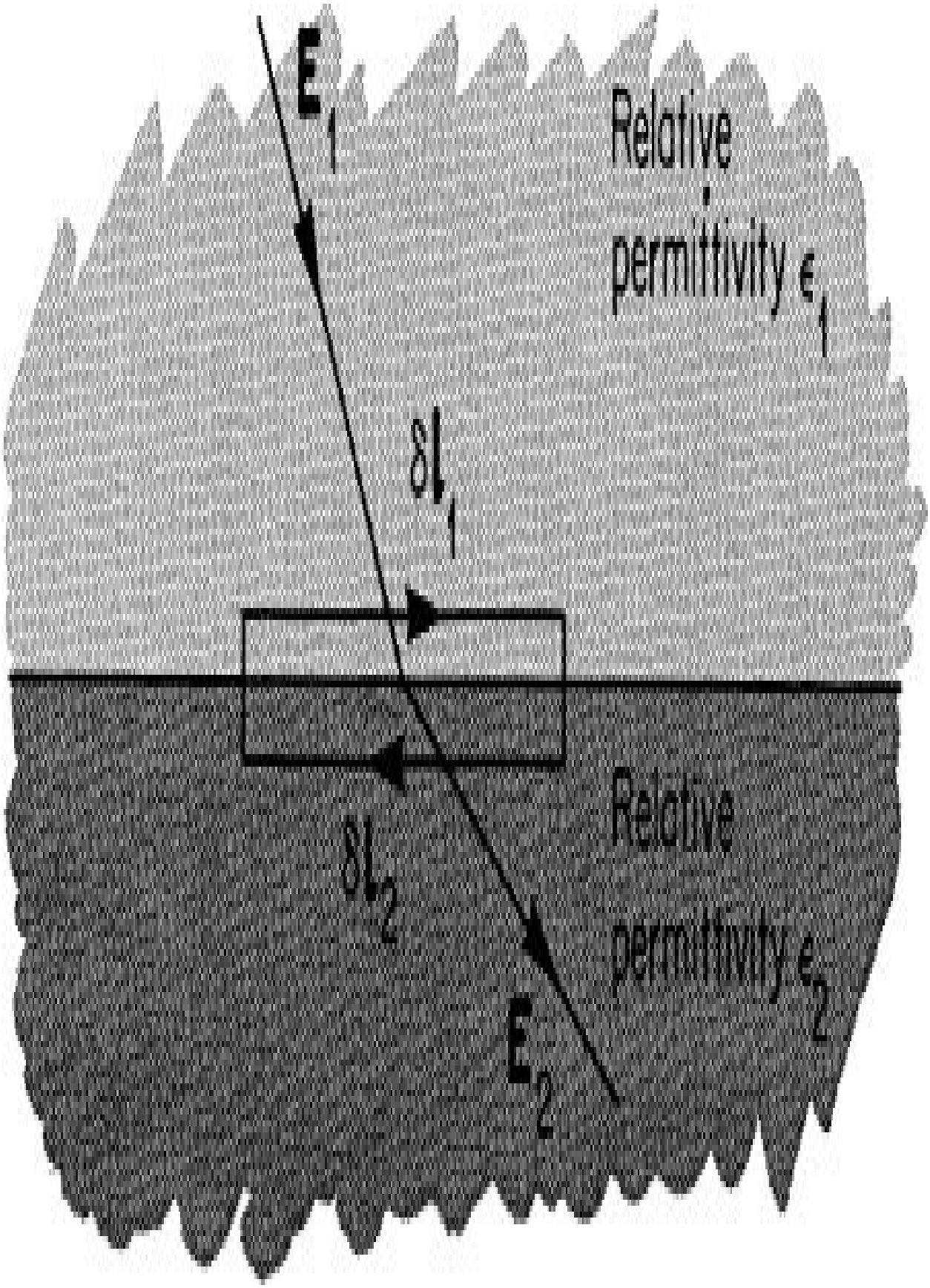
This is true for any δI1 parallel to the boundary, whatever its direction along the surface. It follows that the component of the electric field parallel to the surface is continuous across the boundary. Writing the components as E1|| and E2|| in the two media, we have E1|| = E2||.

Summarizing, at a boundary between two dielectric media which carry no surface charges,

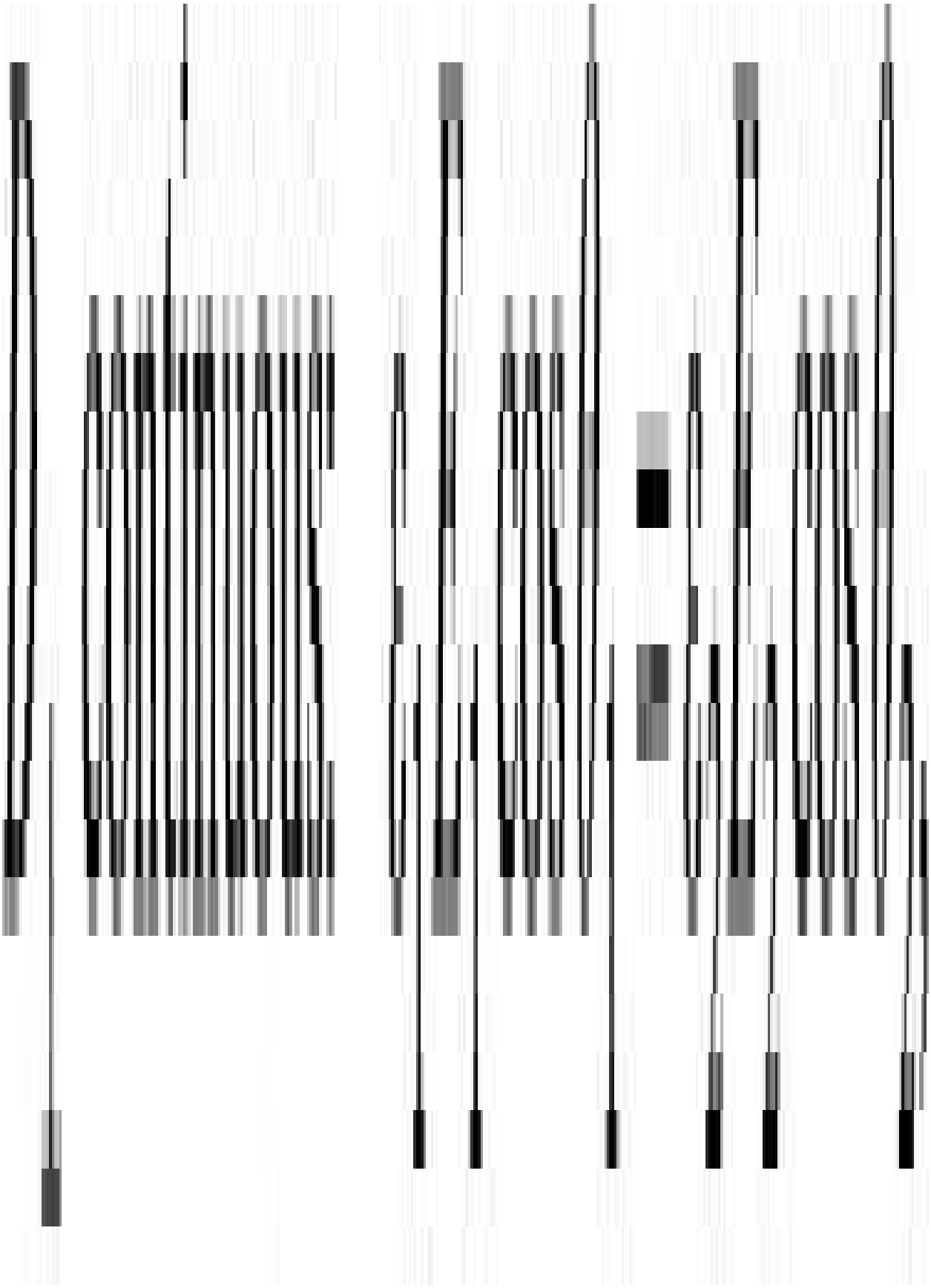
(2.23)

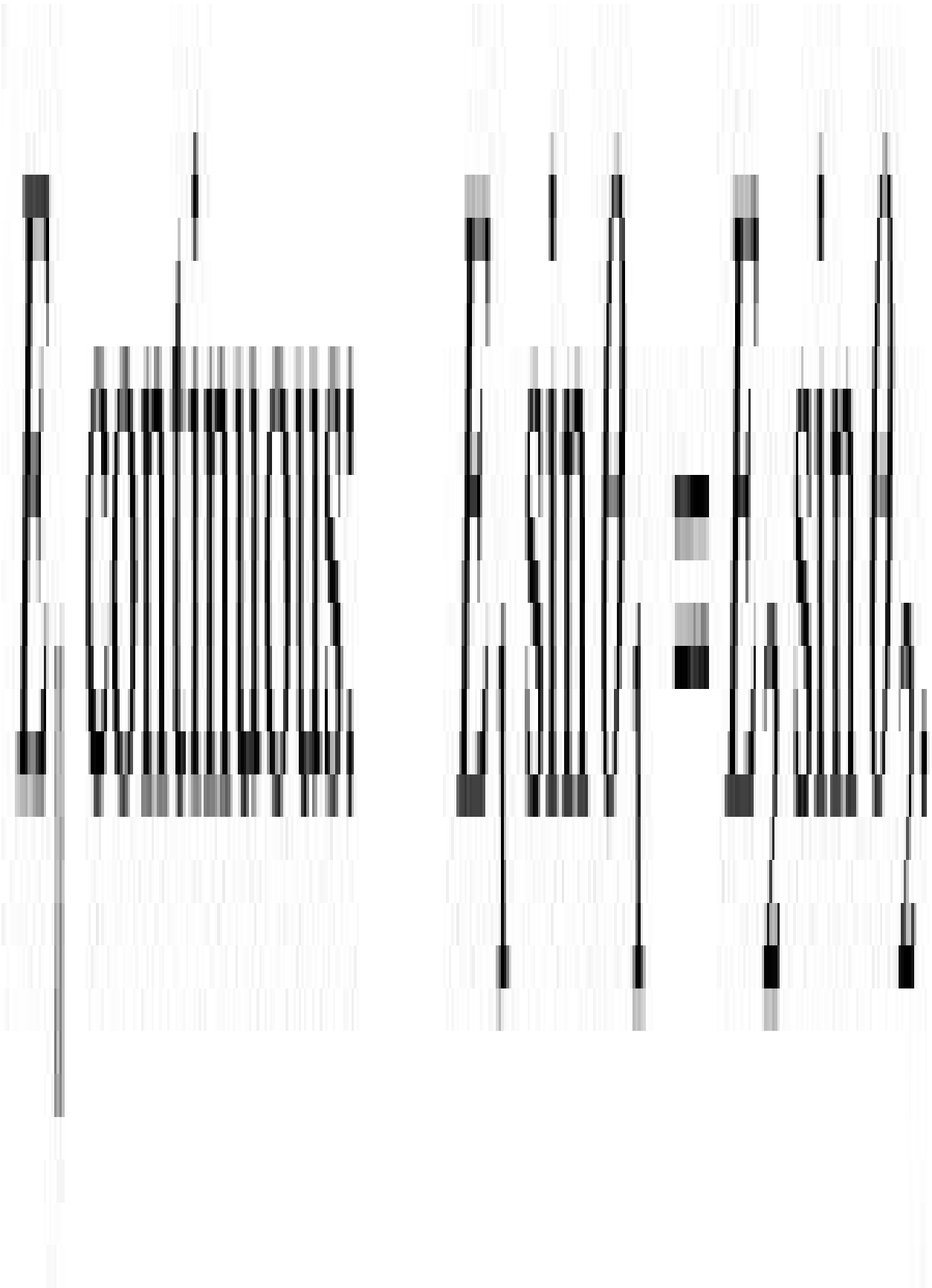
(2.24)

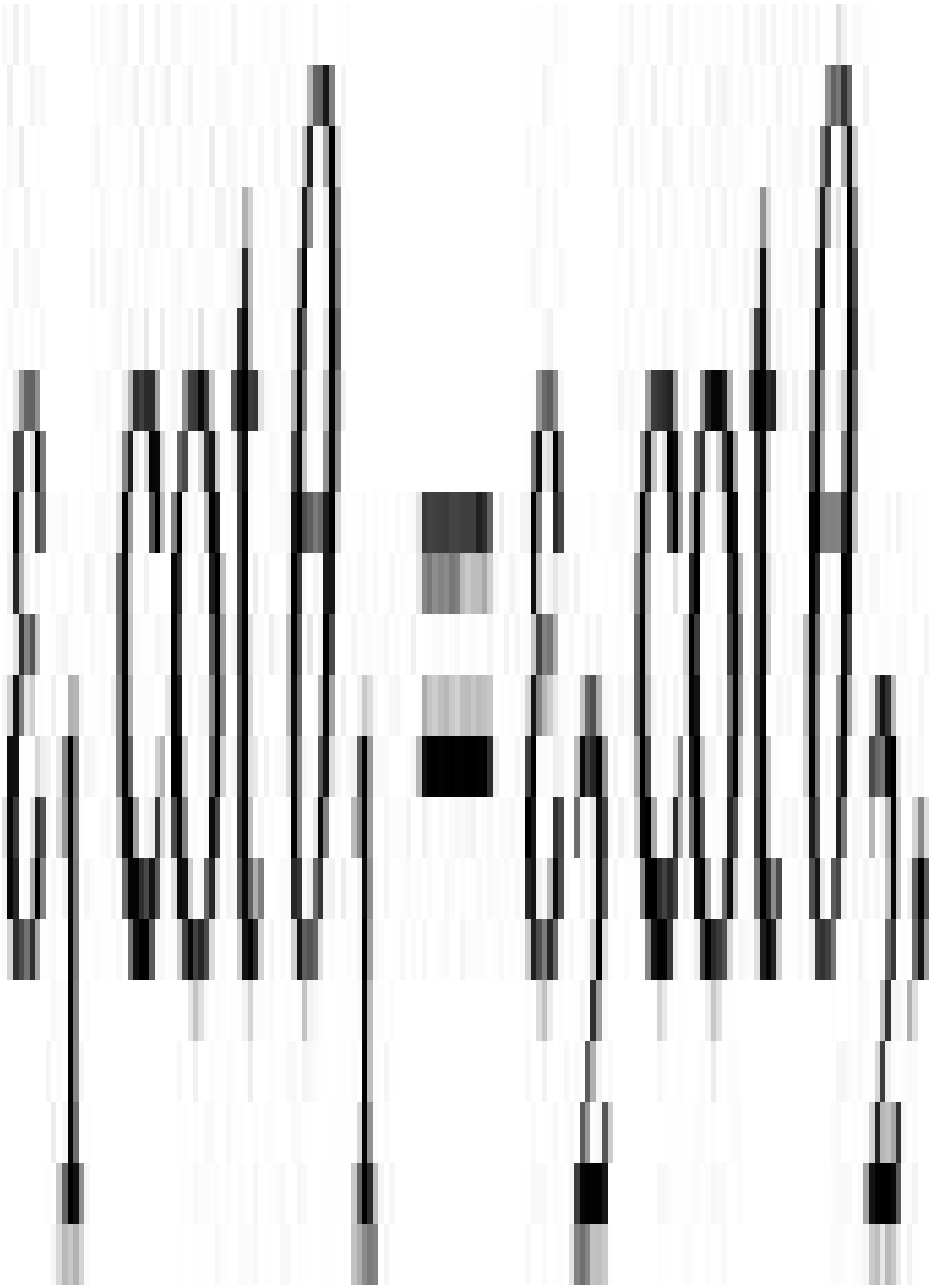
Figure 2.18. The change of E at a dielectric boundary.


E_1
Relative permittivity ϵ_1
δL_1
δL_2
Relative permittivity ϵ_2
E_2

When a boundary is neither parallel nor perpendicular to the electric field, the direction of the field changes across the boundary. Labelling the angles to the normal as in Figure 2.17, the boundary conditions become





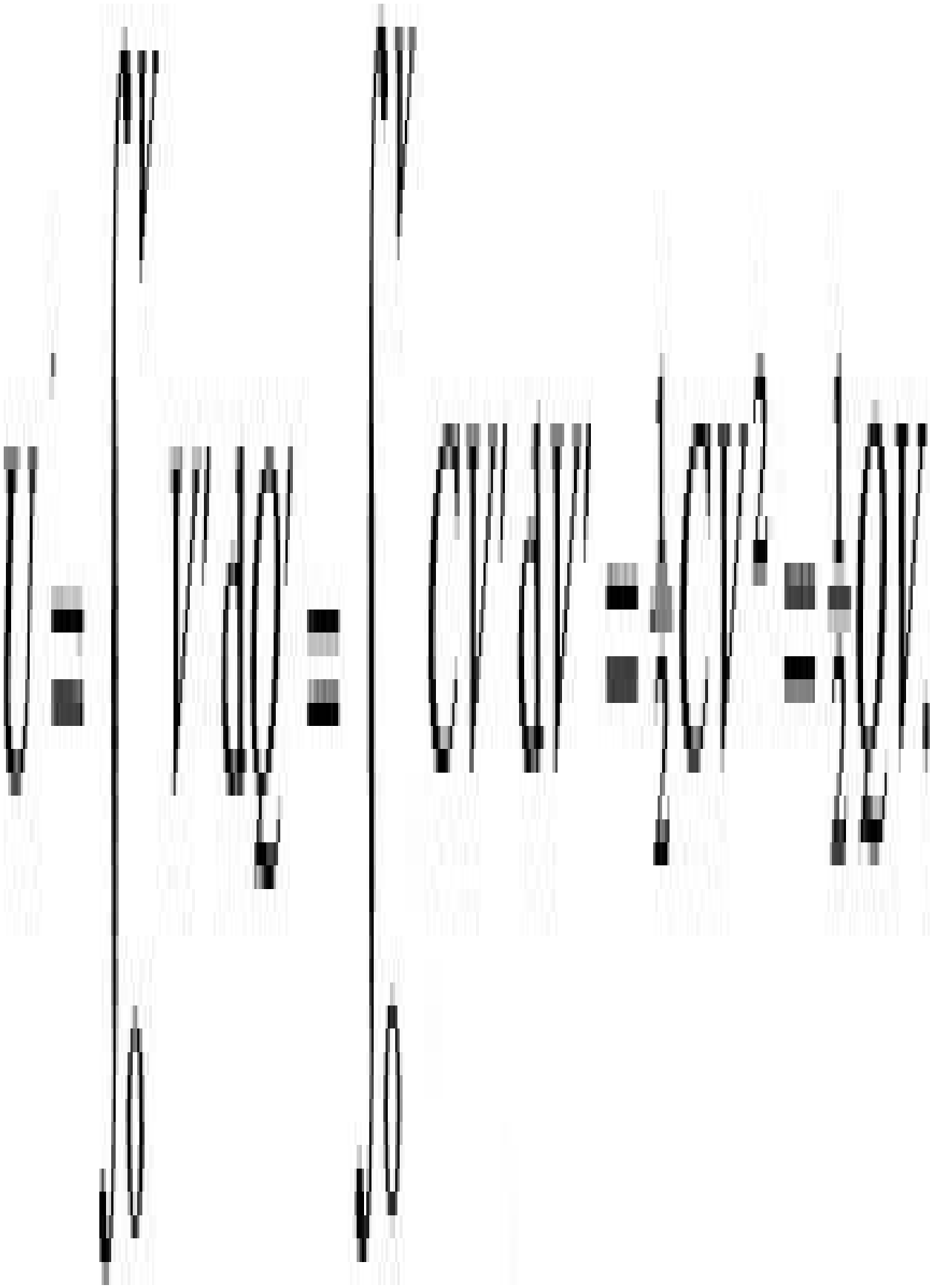


This equation is reminiscent of Snell’s law in optics. The resemblance is no accident. Optical refraction can be explained in terms of the boundary conditions applying to changing electric and magnetic fields; in Chapter 11 we shall derive Snell’s law, and relate refractive index to the relative permittivity.

2.4 ENERGY IN THE PRESENCE OF DIELECTRICS

What is the potential energy of charges in the presence of dielectric material? This is a rather tricky problem, and before dealing with a general system of charges, we shall start by working out the energy stored by a single capacitor. The capacitor is initially uncharged, and a potential difference V is built up between the plates by transferring a charge Q = CV from one to the other. Energy—supplied for example by a battery—is needed to move the charge, and at the moment when the potential difference is V′, energy V′ dQ′ is required to transfer a further charge dQ′. Now dQ′ = CdV′, and the total energy stored is

(2.25)



This expression has the same form as Equation (1.38) for the energy of a vacuum capacitor. The actual value of the energy is of course affected by the presence of dielectric material, since C depends on the relative permittivity. Nevertheless, only Q, the free charge on the capacitor plate, appears in Equation (2.25) which contains no explicit reference to polarization charges. This is because it is only the free charge which is directly moved across the potential difference by external forces.

In a parallel plate capacitor of area A and plate separation d, the energy stored when it is filled with dielectric is $CV^2 = \varepsilon_0\varepsilon(A/d)V^2$. The volume of the capacitor is (Ad), and therefore the energy density is

Just as with fields in vacuum, which were discussed in section 1.5.5, we can regard the energy density DE as residing in the field. As before, it is plausible that the energy density is DE in any electrostatic field, because we can imagine that the field is split up into a large number of parallel plate capacitors by thin conductors placed along closely spaced equipotentials. The total electrostatic energy stored in a volume V is

(2.26)

This result is proved rigorously in the next section.

2.4.1 Some further remarks about energy and forces

To find the energy stored by a general system of charges, we use the same technique as was applied to the capacitor in the previous section. We begin with unpolarized dielectric with no free charges, and then assemble the free charges by bringing them up from infinity. External work is done in bringing up the free charges, which automatically accounts for the potential energy of any polarization charges which may appear. The energy required to assemble a set of charges is (Equation (1.36)), where ϕi is the potential at the position of qi. If the free charges are distributed with surface charge density σf on a number of conducting surfaces, and with volume charge density ρf in the region V bounded by the conductors, the sum is replaced by an integral, and

(2.27)

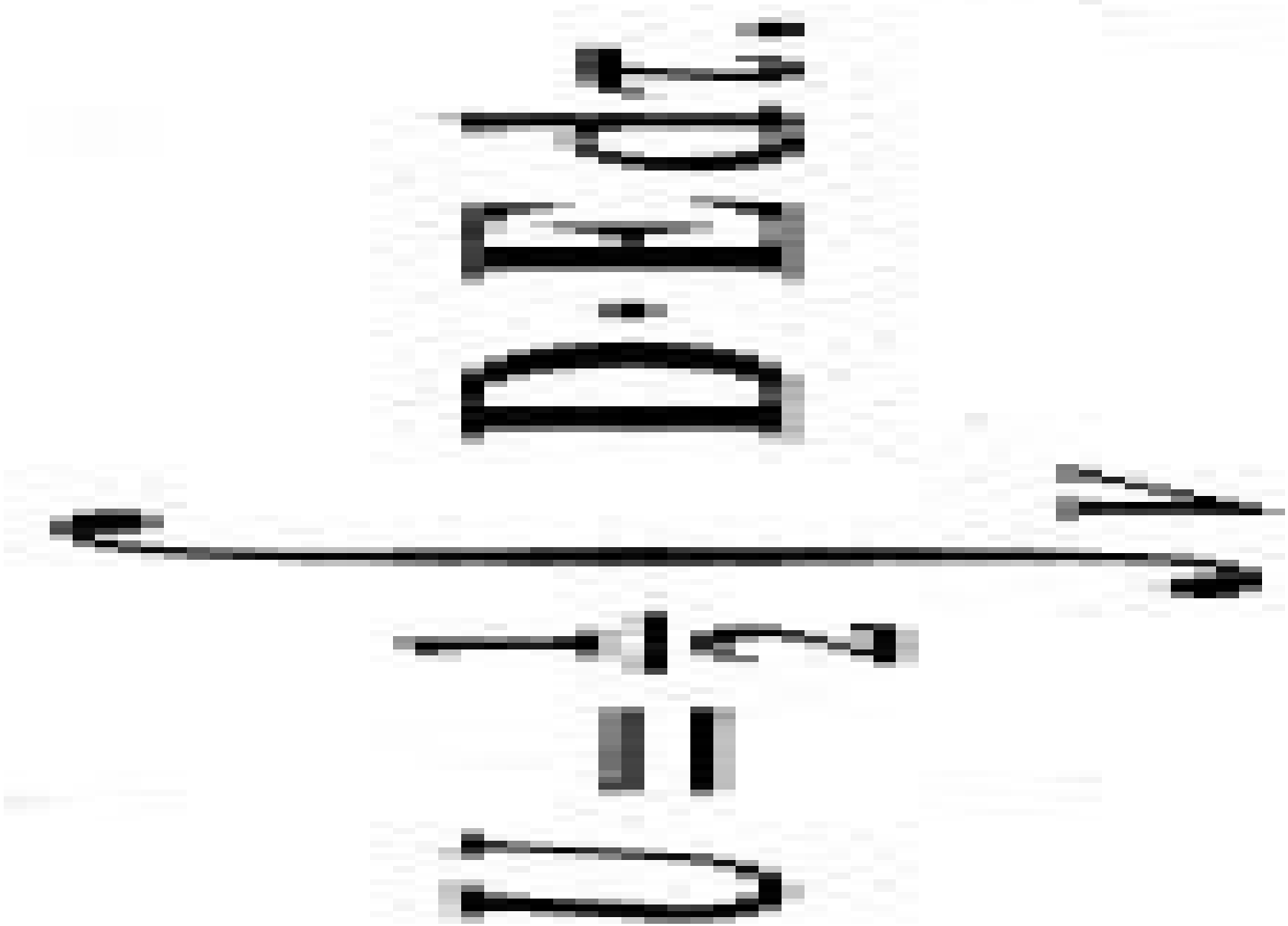
Here, as in the analogous problem of the energy density in vacuum which was treated in section 1.5.6, S includes all conducting surfaces on which there are surface charges and also a distant surface enclosing the whole system.

On the conducting surfaces the outward normal to S is the inward normal to the conductor and by considering the flux of D out of the small box (illustrated in Figure 2.19) on the surface we find $\sigma_f \delta S = -\mathbf{D}\cdot\delta\mathbf{S}$. The volume density of free charge is given by $\rho_f = \text{div } \mathbf{D}$. Substituting in the integrals, the potential energy becomes

$$U = \frac{1}{2}\int_V \phi \operatorname{div} \mathbf{D}\, d\tau - \frac{1}{2}\int_S \phi \mathbf{D}\cdot d\mathbf{S}$$

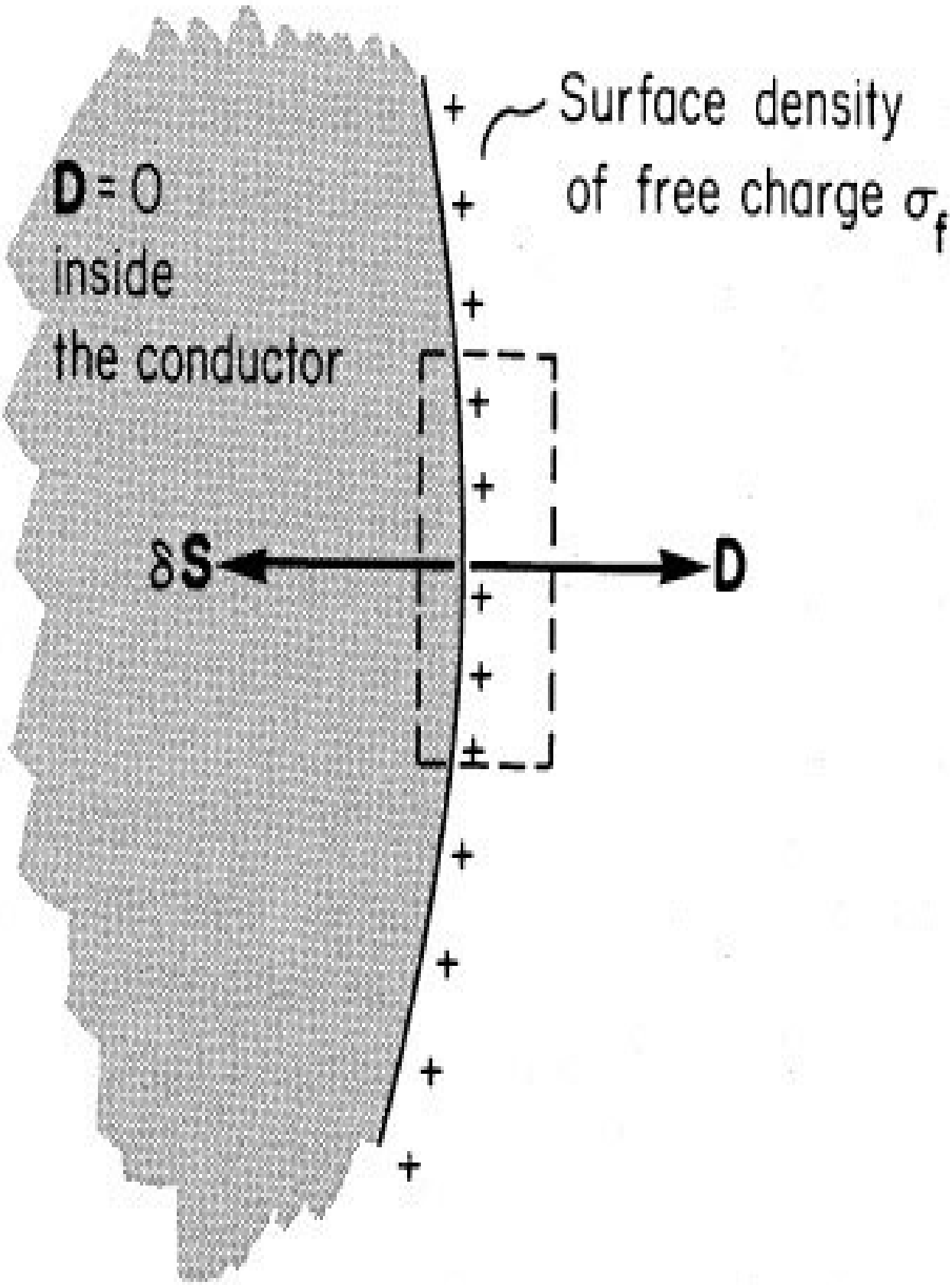
$$= \frac{1}{2}\int_V \phi \operatorname{div} \mathbf{D}\, d\tau - \frac{1}{2}\int_V \operatorname{div}(\phi \mathbf{D})\, d\tau,$$

Using the vector identity div (ϕD) ≡ ϕ D + D·grad ϕ, and putting E = − grad ϕ, we find that the potential energy is indeed given by Equation (2.26):



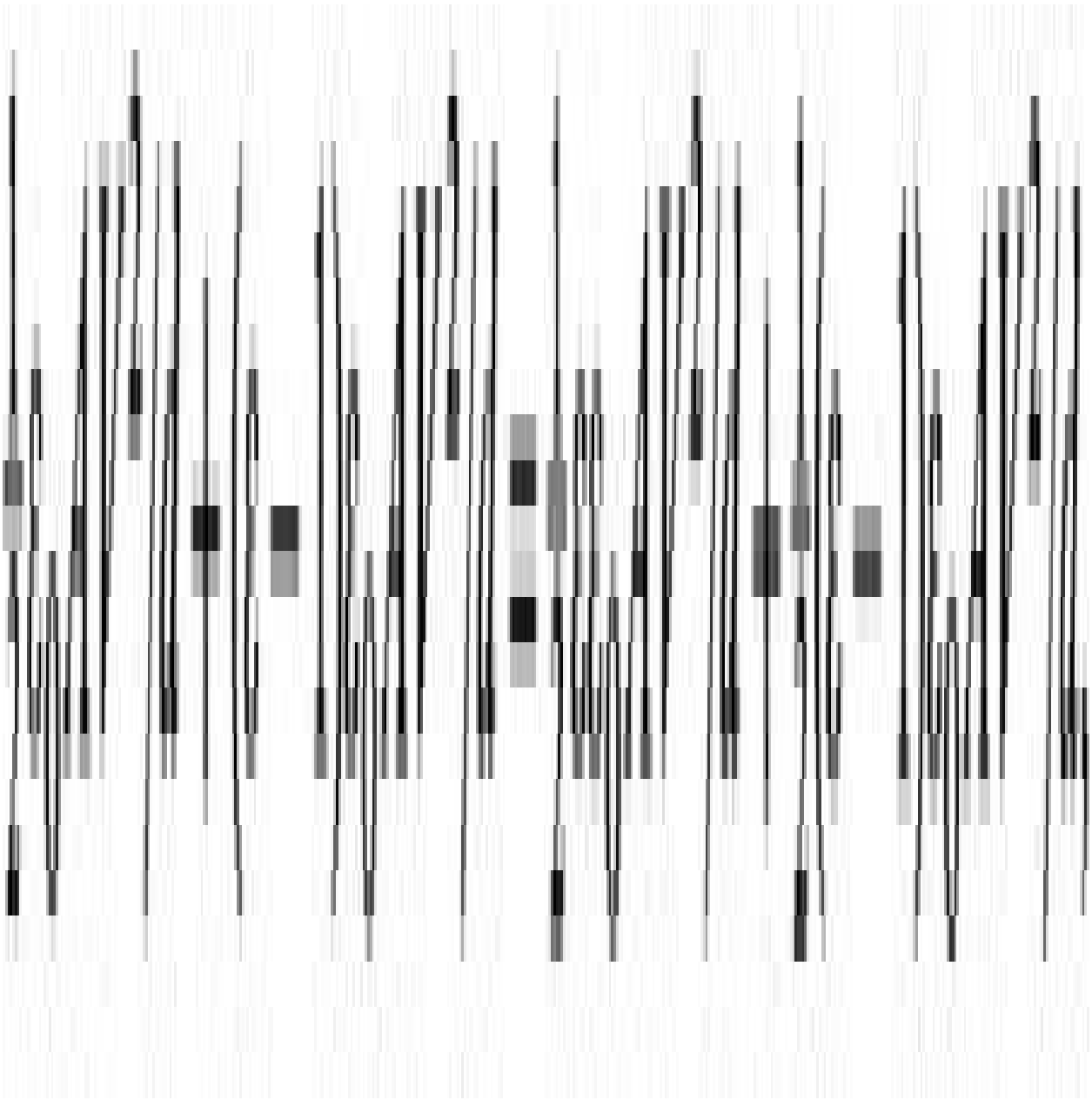
A word of caution must be added here. In assembling the system of charges piece by piece to arrive at Equation (2.27), we have tacitly assumed the polarization to be proportional to the electric field. Except in some very uncommon materials this is always true, but where non-linear effects do occur, the external work done may depend on the precise path by which the final state of the system is approached. Energy is of course always conserved, but sometimes it is not possible to make a unique distinction between potential energy and heat. We shall meet the equivalent problem in magnetism, where non-linear materials are of practical importance.

Figure 2.19. At a conducting surface, D is related to the surface density of free charge.


Surface density
of free charge σ_f
D = 0
inside
the conductor
δS
D

Even with materials in which polarization is proportional to the field, when calculating the potential energy one must be careful to follow the recipe given above, and assemble the free charges when the matter is already in place. Suppose that we reverse the order for the parallel plate capacitor, first charging it up, and then afterwards inserting the dielectric while maintaining the potential difference constant across the plates. In vacuo an amount of work is done in raising the potential difference across the plates to V. When the dielectric is inserted, an additional charge $(\varepsilon - 1)\varepsilon_0 AV/d$ must be added to keep the potential difference at the value V, requiring additional work $(\varepsilon - 1)\varepsilon_0 AV^2/d$.

The total amount of external work is



more than the amount $\varepsilon\varepsilon_0 AV^2/d$ expended when the dielectric was in place before the free charges were moved. The difference $(\varepsilon - 1)\varepsilon_0 AV^2/d$ occurs because there is a force on the dielectric, and work is done on it as it enters the capacitor. To be definite, let us assume that the capacitor plates are square and of side a. When the dielectric is inserted a distance x into the capacitor, extra work $(\varepsilon - 1)(\varepsilon_0 aV^2/d)\,dx$ is done on the dielectric if it is moved a further distance dx, and the force is therefore $(\varepsilon - 1)\varepsilon_0 aV^2/d$. The force is independent of x because it only acts at the edge of the plates, where the field is non-uniform. The same argument applied to a general field shows that the force per unit volume acting on a piece of dielectric is grad $(\varepsilon_0 E^2 - \mathbf{D}\cdot\mathbf{E})$. This is a force which attracts dielectric material towards regions where the electric field is high.

PROBLEMS 2

2.1 Two identical capacitors are connected in parallel charged by a battery of voltage V, and then isolated from the battery. One of the capacitors is then filled with a material of relative permittivity ε. What is the final potential difference across its plates?

2.2 A capacitor is placed in a tank, and the following sequence of events occurs:

(i) a battery of voltage V is connected across the capacitor terminals;

(ii) the tank is filled with oil of relative permittivity ε;

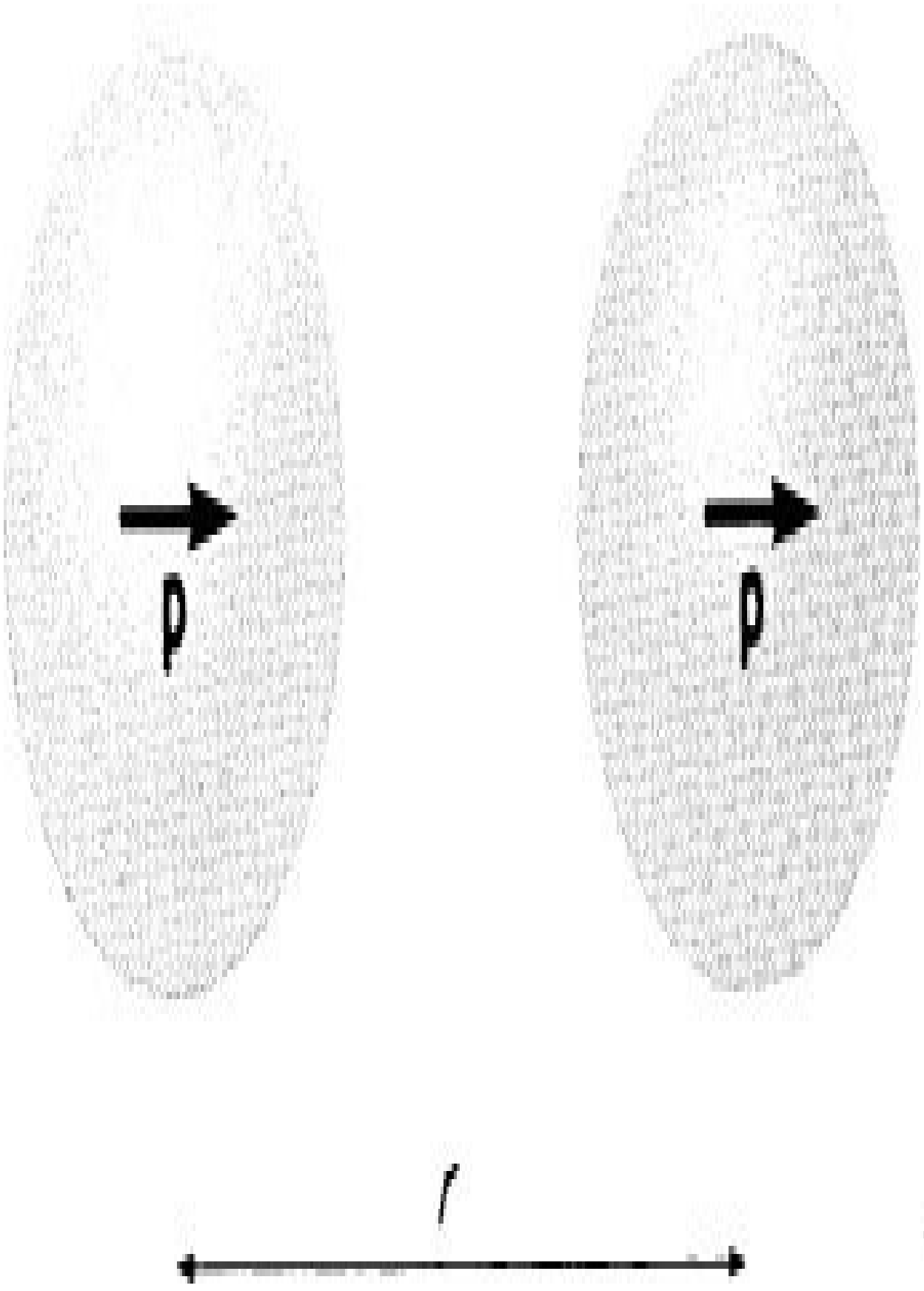
(iii) the battery is disconnected, and the tank is afterwards drained.

What is the energy stored by the capacitor at each stage? How do you reconcile the energy changes with conservation of energy?

2.3 A 1 μF capacitor is made up of thin sheets of polycarbonate (the transparent foil used as a wrap in cooking) which have a conducting film deposited on them. The foil is rolled up to form a cylinder of diameter 1 cm and length 2 cm. The relative permittivity of polycarbonate is 2.3. Estimate the thickness of the polycarbonate sheet. At the maximum recommended voltage of 50 volts across the terminals of the capacitor, what is the force per unit area compressing the dielectric?

2.4 At room temperature the relative permittivity of water is 80. The dipole moment of a water molecule is 6.2×10^{-30} coulomb metres. What is the average value of Elocal/E for a water molecule? (In working out this problem, neglect the contribution to the relative permittivity from induced dipole moments.)

Figure 2.20.


p
p
r

2.5 Two molecules each have a dipole moment p pointing along the line joining their centres, as shown in Figure 2.20. How does the force between the molecules vary with their separation r? If two water molecules are oriented in this way, what is the potential energy due to their dipole–dipole interaction when their centres are at their average separation of 3.1×10^{-10} m? (Dipole moment of a water molecule = 6.2×10^{-30} coulomb m.)

2.6 Imagine that the dipoles in Figure 2.20 have their centres fixed, but are both free to rotate and take up any orientation. Think about the couples the dipole moments can exert on one another, and sketch the orientations in which they are in stable or unstable equilibrium.

2.7 A slab of dielectric of relative permittivity ε fills the space between $z = \pm a$ in the x–y plane, and contains a uniform density of free charge ρ_f per unit volume. Find E, D and P as functions of z. What is the surface density of polarization charge on the surface of the dielectric?

2.8 A slab of dielectric of thickness a and relative permittivity ε is placed in a uniform external field E whose field lines make an angle θ with normals to the surface of the slab. What is the density of polarization charge on the surface of the slab? Neglect end effects.

* The magnitude and direction of the induced dipole moments in some crystalline materials depend on the orientation of the crystal axes with respect to the field direction. Such materials are called anisotropic dielectrics, in contrast to isotropic dielectrics which have the same response to an electric field whatever their orientation, and for which induced dipole moments always point along the field direction. In this book we shall only consider isotropic dielectrics.

* The relative permittivity is often called the dielectric constant, especially in books which do not use SI units.

* This value applies to water in a static field. The relative permittivity is smaller

in rapidly changing fields: this phenomenon is discussed in section 11.2.2.

* The rôle of the Boltzmann factor in equilibrium probability distributions is discussed in section 2.5 of the Manchester Physics series volume on Statistical Physics, by F. Mandl, 2nd edition, Wiley, 1988.

* It is not rigorously correct to replace $Px(x + \delta x) - Px(x)$ by $(\partial Px/\partial x)\,\delta x$, because δx can never be allowed to tend to zero—it must always be large compared with atomic sizes. There are corrections to Equation (2.17) which are small if the fractional change in polarization is small over a distance of one atomic spacing. In practice these corrections are completely negligible.

CHAPTER 3

Electrostatic field calculations

Information about the electric field may be required in a variety of different physical situations. For example, we may want to know where breakdown is most likely to occur in a piece of high-voltage equipment; or what path an electron follows as it travels through an electron microscope; or the value of the capacitance between a pair of conductors. To find the answer to each of these problems, we must first work out the electric field. Often the electric field cannot be expressed in terms of a finite number of elementary functions, and the best that can be done then is to look for approximations. Even when a manageable solution to the problem exists, there is no universal way of finding out what it is. One must know (or guess) the functional form of the solution, and fit it to the problem in hand, in much the same way as one guesses the result of integrating a function by comparison with standard forms of integral, and then checks by differentiation. Some of the techniques which can be used to lead to solutions of electrostatic problems are outlined in this chapter. The examples discussed are not simply intended as mathematical exercises, but have been chosen to illustrate important properties of the electric field.

When fields have a very simple symmetry, they can often be calculated directly from the integral form of Gauss' law (Equation (1.14)). For example, the electric displacement between the conductors of a coaxial cable has already been found in this way. This was possible because the cylindrical symmetry of the cable itself implies that the displacement D is also cylindrically symmetrical, that is to say, that it depends only on the radial distance r from the axis of the cable. Consideration of the flux of D out of a cylindrical surface then shows immediately that D is proportional to 1/r, as we found in section 2.3.2. Since the dielectric in the cable is uniform, the electric field E is also proportional to 1/r.

The same procedure works for problems with plane and spherical symmetry. One imposes the limitation that D has the appropriate symmetry, and then applies Gauss' law over a surface on which D is constant. We have already used this technique to find the relation between electric field and charge in a parallel plate capacitor and to find the atomic field around a spherically symmetrical ion.

3.1 POISSON'S EQUATION AND LAPLACE'S EQUATION

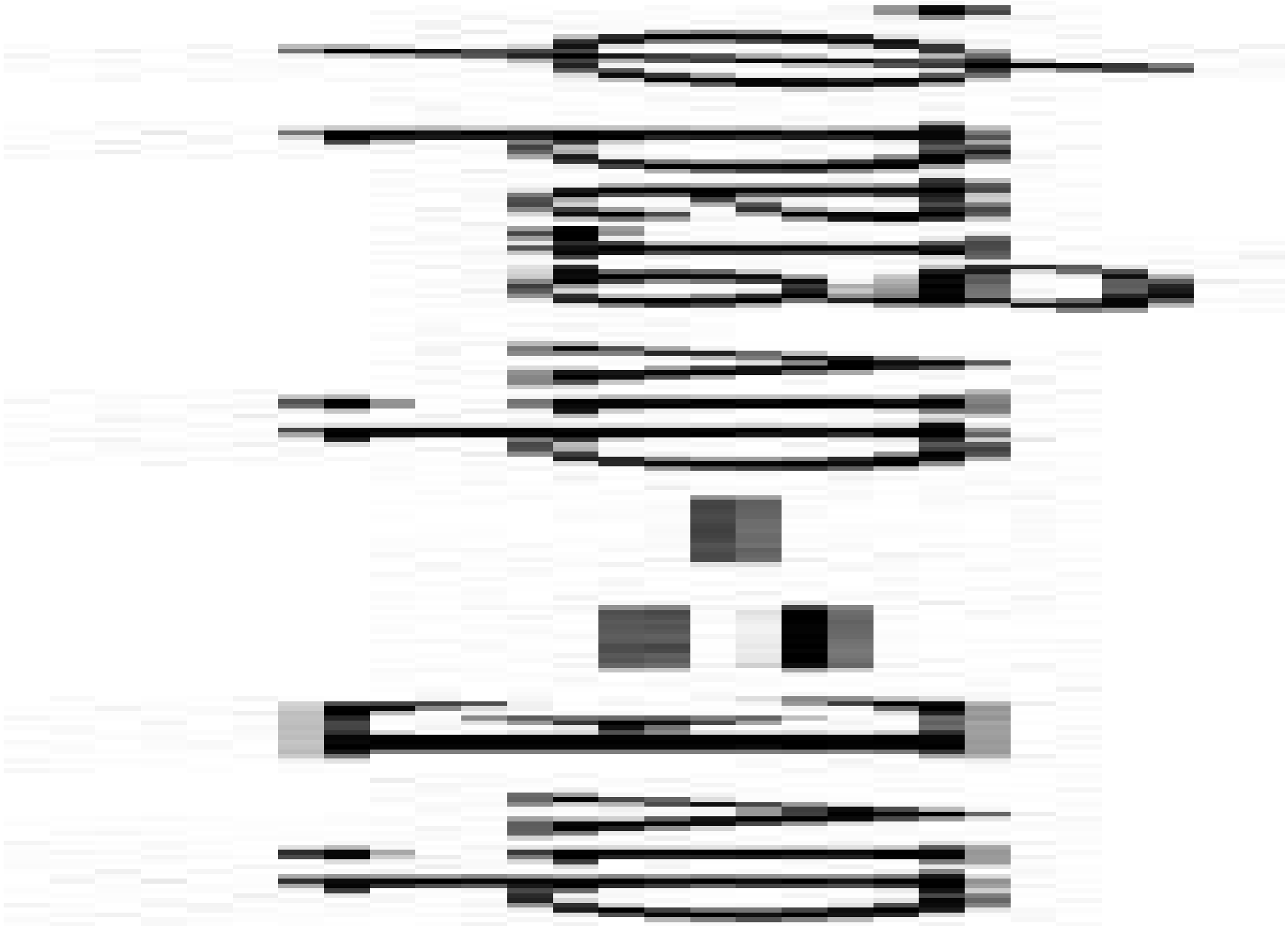
In most practical problems in which the electric field needs to be known, there is no simple symmetry. It is then usually better to begin with the differential form of Gauss' law rather than the integral form. We shall discuss the differential form in this section, illustrating its properties with very simple examples.

Let us start by applying Gauss' law in its differential form to the coaxial cable, in which we already know the electric field. Assume that the dielectric material between the conductors is uniform, and has relative permittivity ε. The outer conductor, of radius b is at zero potential, and the inner conductor, of radius a, is held at a potential V. We write Gauss' law in its differential form (Equation (1.20))

(3.1)

where ρ is the total charge density, including any polarization charges.

The electric field must obey this equation everywhere between the conductors, and it must also satisfy the boundary conditions of the problem. The boundary conditions are that the potential is constant over the surface of the conductors, and has the values 0 and V on the outer and inner conductors respectively. Because the boundary conditions refer to the potential, it is convenient to express Gauss' law in terms of the potential as well. The field is related to the potential by $E = -\text{grad}\ \phi$, and hence



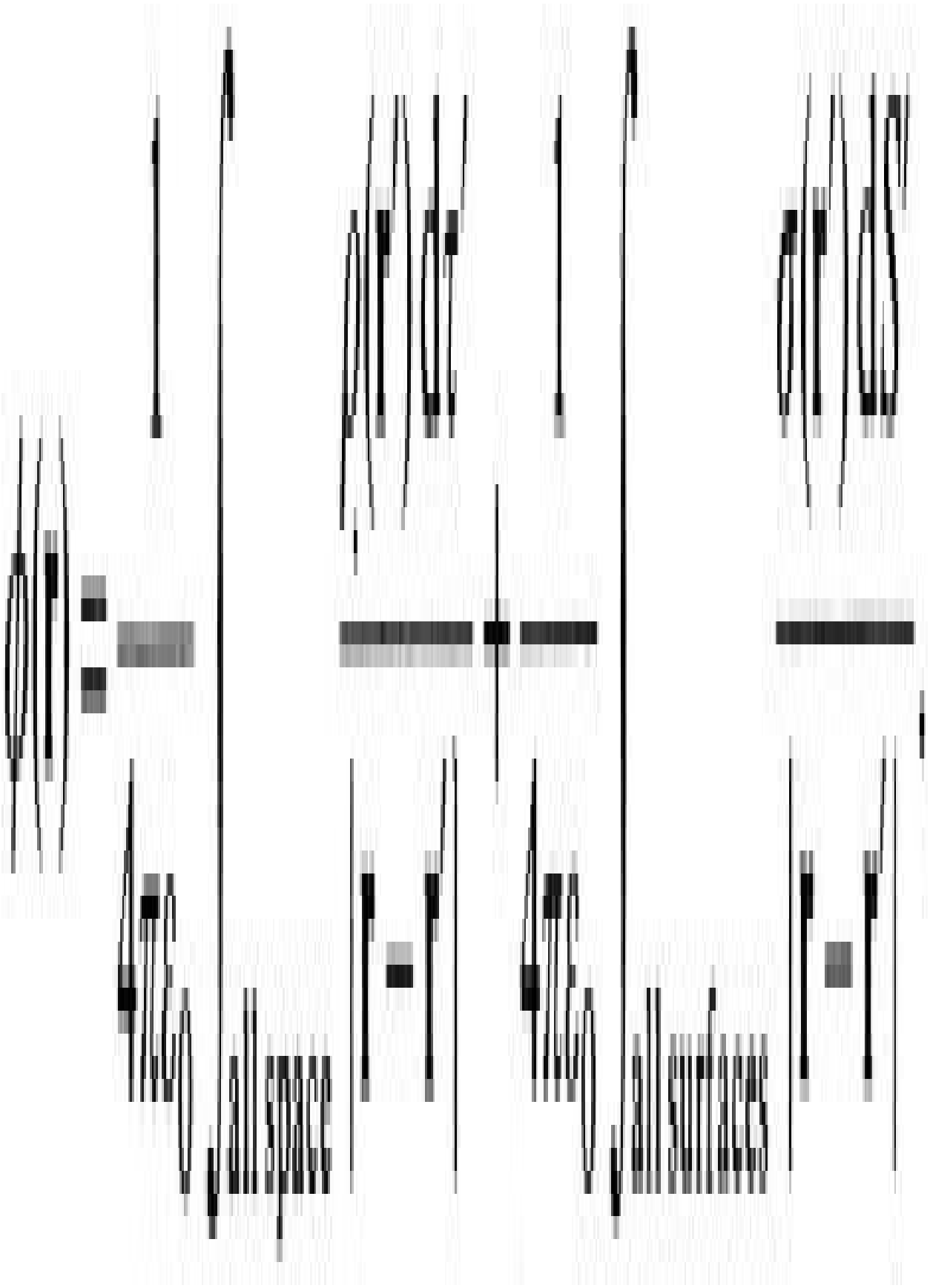
The combination (div grad) is called the Laplacian operator. It is a second order differential operator, and it is usually written in concise form as ∇^2. Now the gradient (grad ϕ) of a scalar function of position ϕ is the vector with components $(\partial\phi/\partial x, \partial\phi/\partial y, \partial\phi/\partial z)$ in Cartesian coordinates. The divergence of this vector is $(\partial^2\phi/\partial x^2 + \partial^2\phi/\partial y^2 + \partial^2\phi/\partial z^2)$, and thus in Cartesian coordinates the Laplacian operator is div grad $\equiv \nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$. The forms of the Laplacian operator in cylindrical and spherical polar coordinates are given in Appendix B.

Substituting (− div grad ϕ) for div E in Equation (3.1),

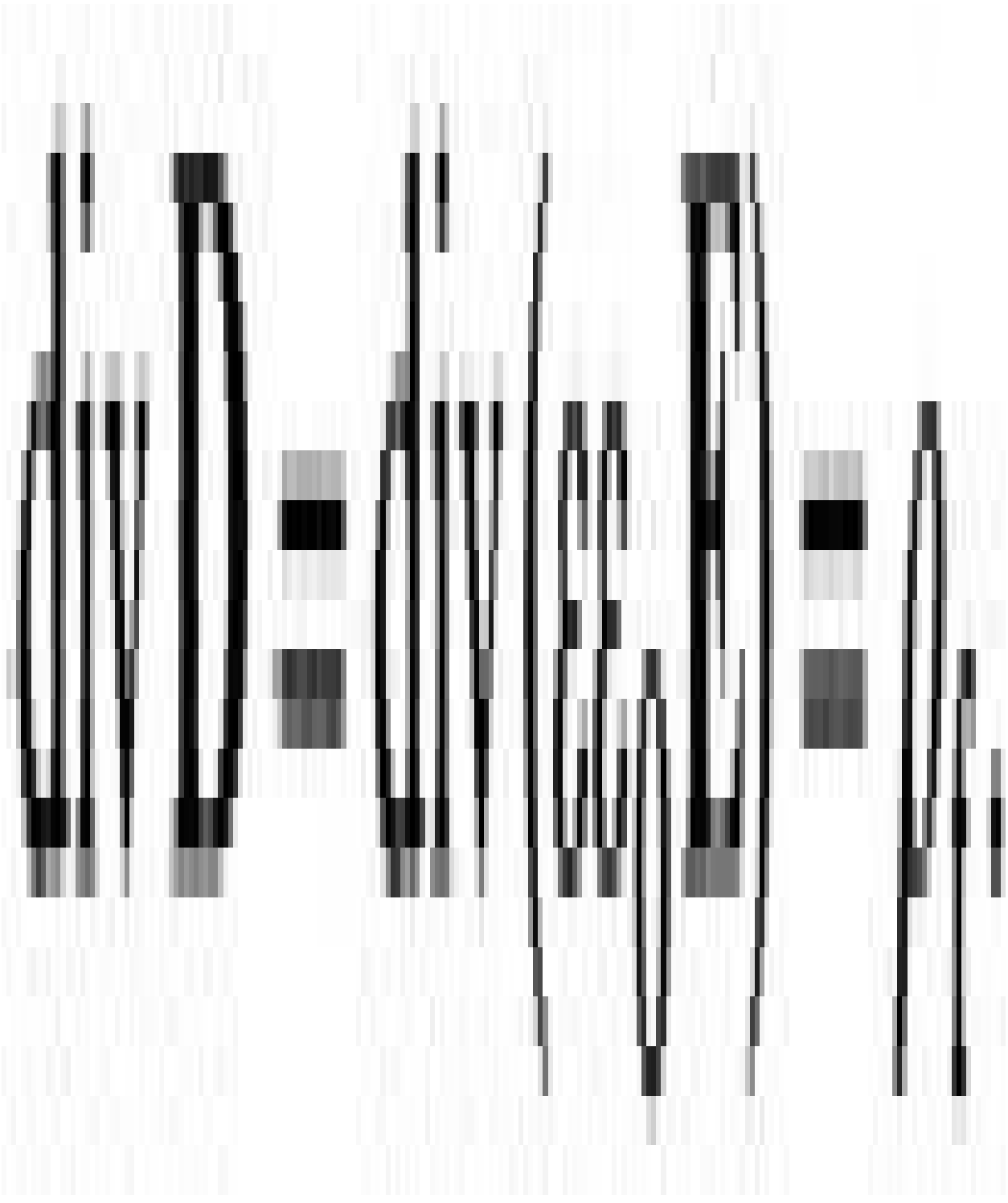
(3.2)

This equation is known as Poisson’s equation.

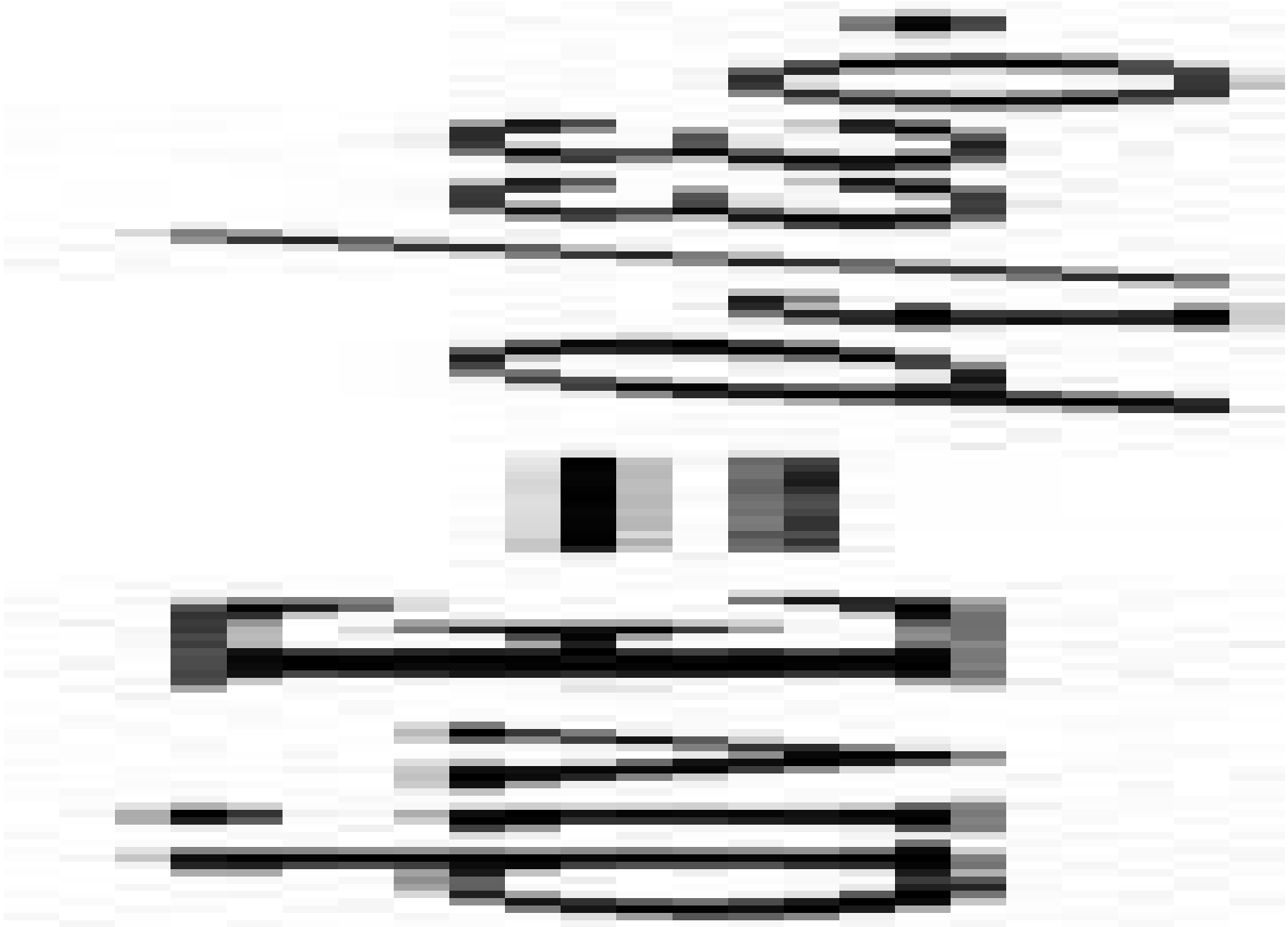
The problem of finding the electrostatic field has now been reduced to the solution of Poisson’s equation, a second order differential equation. A formal solution to Poisson’s equation has already been written down in section 1.5.2, where the principle of superposition was invoked to show that



Here the potential at the point with position vector r is given in terms of the volume charge density ρ and surface charge density σ at the point with position vector r′. This expression for the potential is quite general, and therefore it is a solution to Poisson's equation. But it is not of much use when dielectrics and conducting materials are present, because induced surface charges and polarization charges then appear, of a magnitude which can only be evaluated when the field or potential is already known. The polarization charges must be accounted for by introducing the electric displacement D, and writing Gauss' law in terms of the free charge density as in Equation (2.19):



In a region filled with a uniform dielectric of constant relative permittivity, as in the space between the conductors in the coaxial cable, this equation simplifies to



Again E can be replaced by (– grad ϕ), leading to

(3.3)

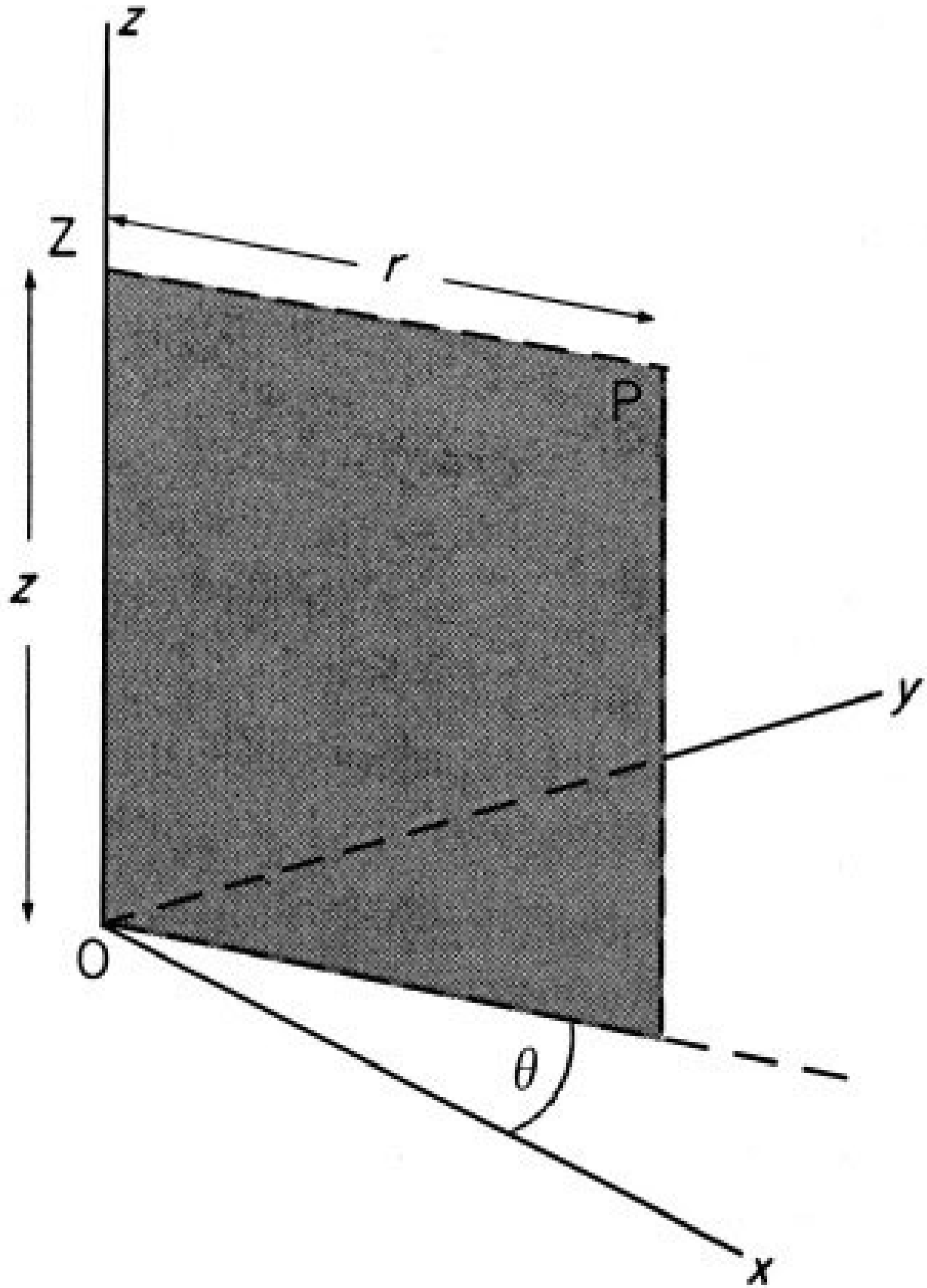
an equation still retaining the form of Poisson's equation. This equation is the one we usually want to use for solving electrostatic problems, since it makes no reference to polarization charges or to induced charges on conductors—the presence of conductors only imposes the condition that the potential ϕ must be constant on any conducting surface.

In the coaxial cable there is no free charge between the inner and outer conductors, and Equation (3.3) simplifies to

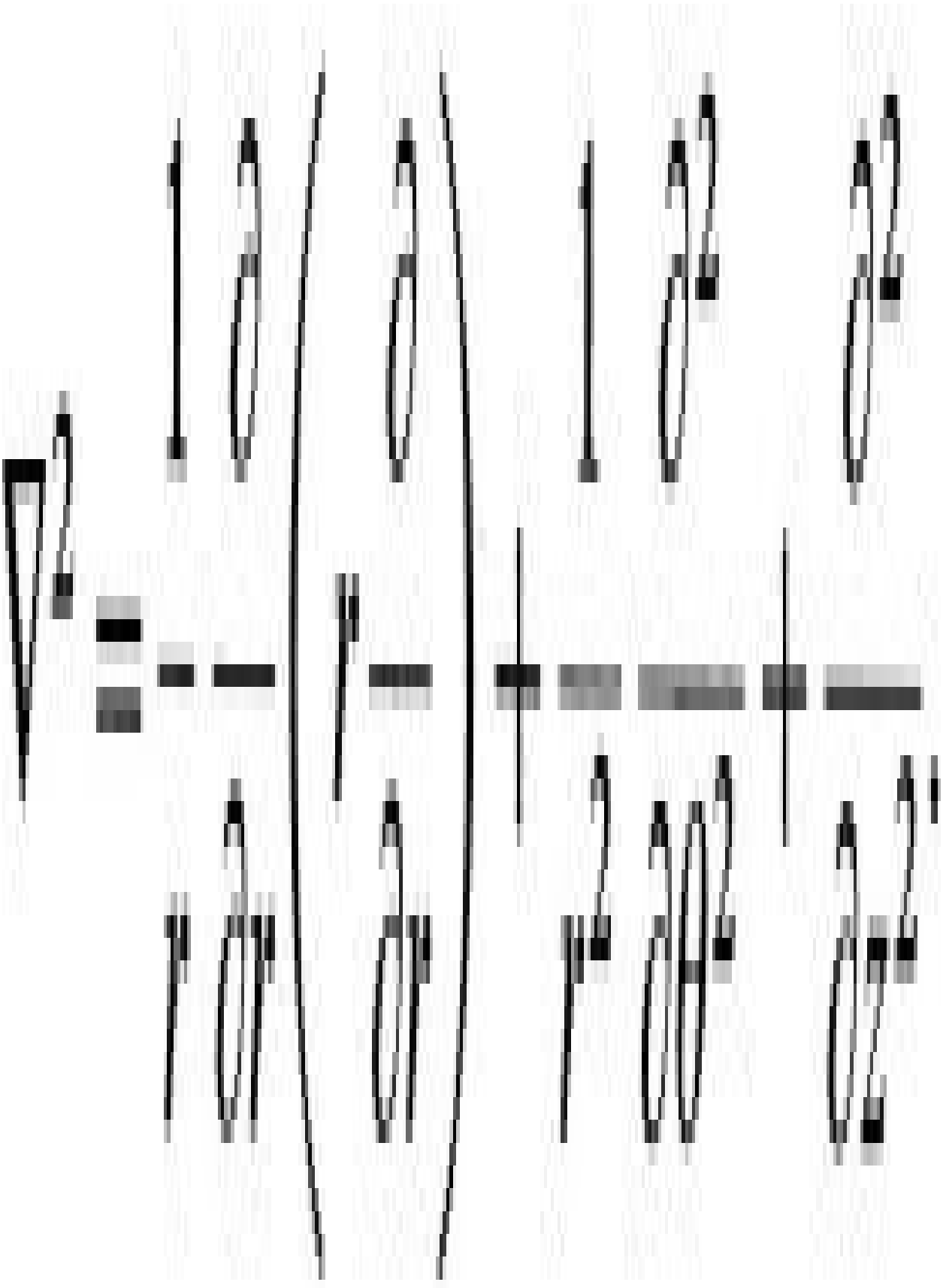
(3.4)

This important special case of Poisson's equation is called Laplace's equation. The appropriate coordinates for the cable are the cylindrical polar coordinates (r, θ, z). In cylindrical polar coordinates r is the perpendicular distance from (r, θ, z) to the z-axis, as shown in Figure 3.1. (The variable r must not be confused with the distance to the origin, which we label r in spherical polar coordinates.) The angle θ is the azimuthal angle of the position vector, with respect to an x-axis drawn through the origin. In cylindrical polar coordinates the Laplacian

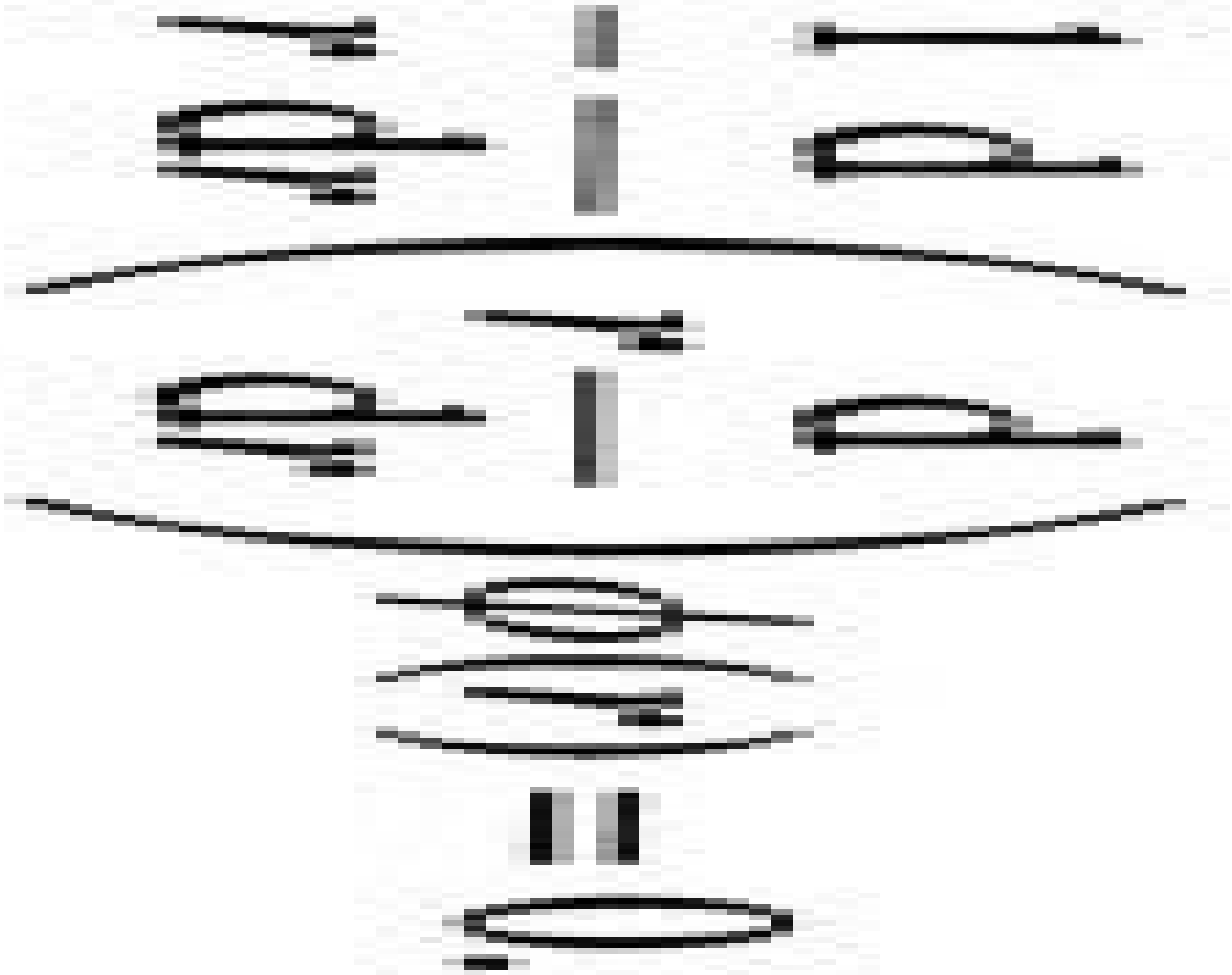
Figure 3.1. Cylindrical polar coordinates.


z
Z
r
P
z
y
O
θ
x

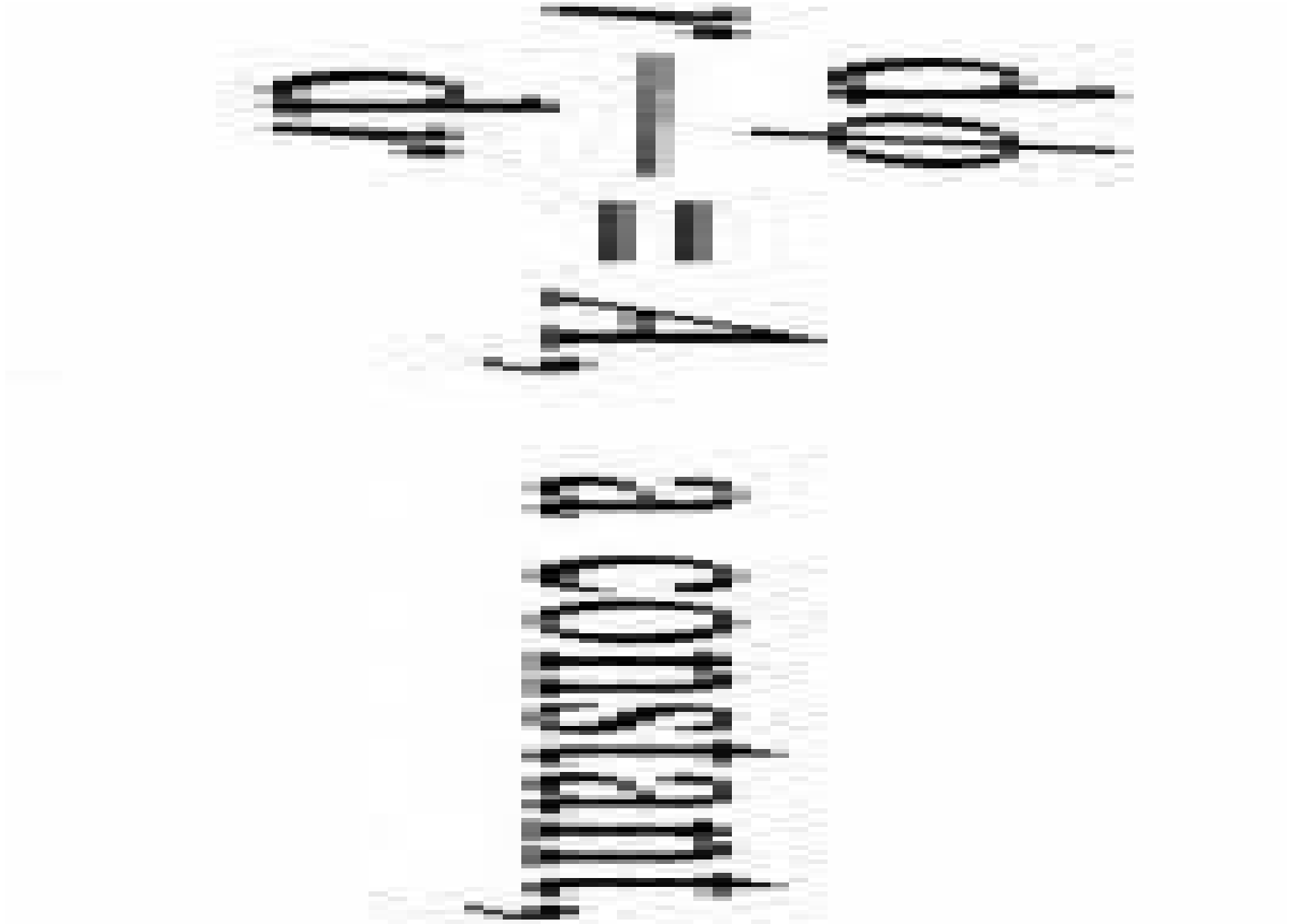
operator can be shown to have the form



Since the coaxial cable has cylindrical symmetry, there is no θ-dependence, and for a long cable we may neglect end effects, so that there is no z-dependence either. Laplace’s equation then reduces to



Hence



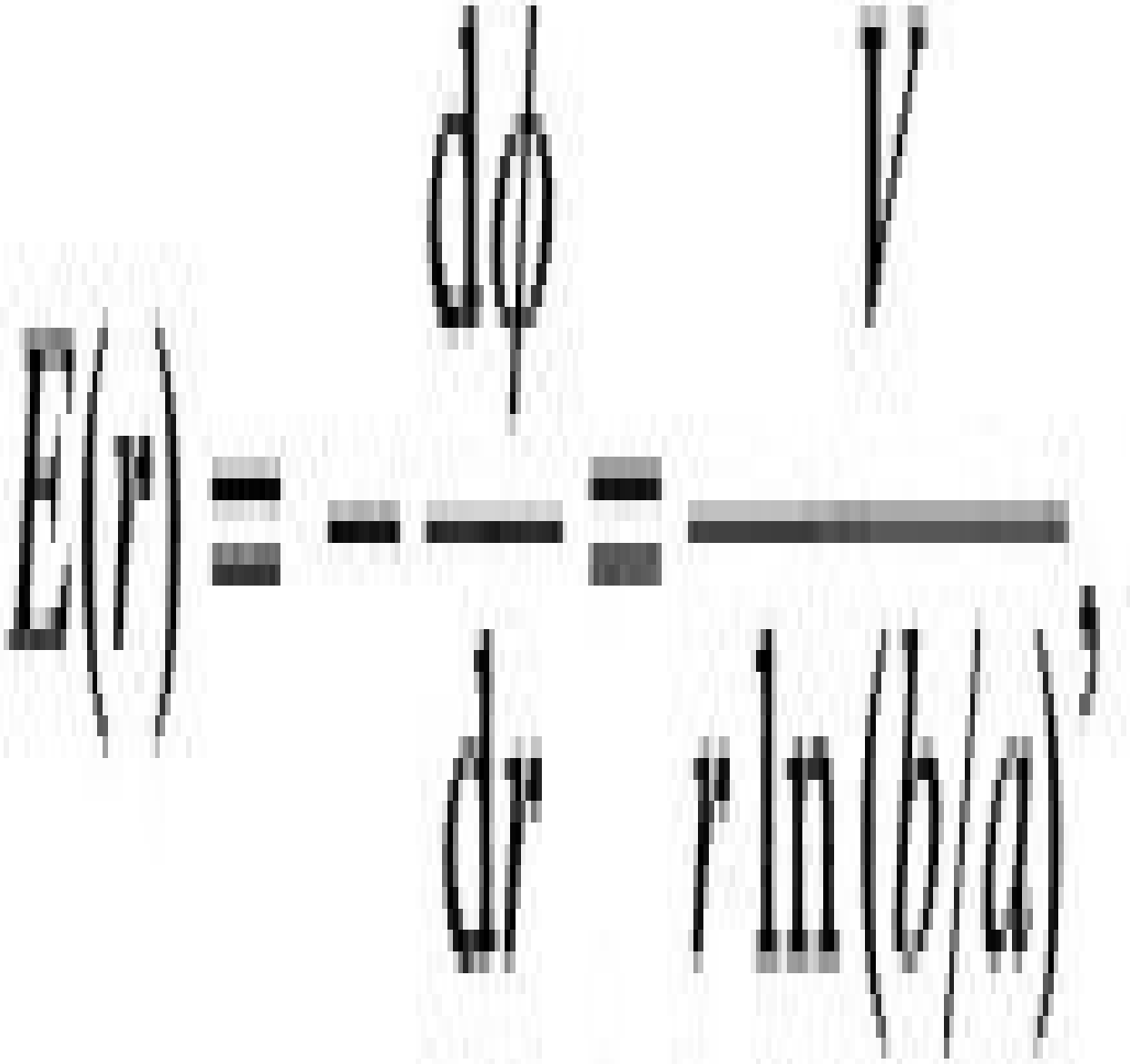
and

$\phi(r) = A \ln(r) + B$, where B is another constant of integration.

Choosing the zero of potential to be at the radius b of the outer conductor, the boundary conditions when a potential V is maintained across the cable are $\phi(a) = V$ and $\phi(b) = 0$, where a is the radius of the inner conductor. The values of the constants of integration A and B which satisfy these conditions lead to

(3.5)

The field has only a radial component, of magnitude

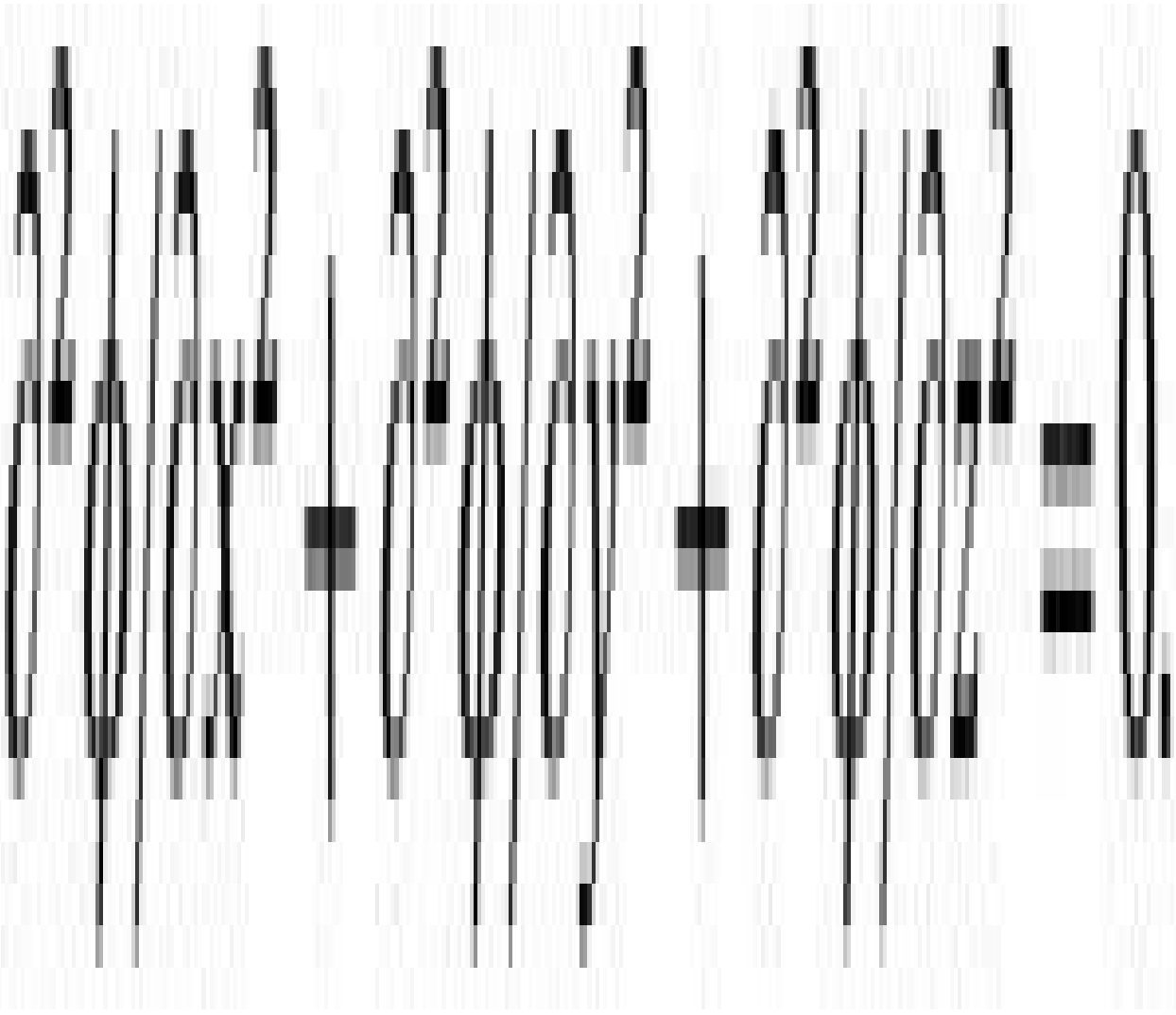


in agreement with the result found in section 2.3.2 from the integral form of Gauss' law.

3.1.1 The uniqueness theorem

The potential $\phi(r)$ given in Equation (3.5) obeys Laplace's equation in the region $a < r < b$ and also satisfies the conditions imposed at the boundaries of this region. Obviously $\phi(r)$ is the only function which has these properties, because we found it by integrating Laplace's equation step by step in a systematic manner. Unfortunately, systematic integration is not generally possible. However, suppose that in some region we have managed to find a solution to Poisson's equation or Laplace's equation, having specified values on the boundaries of the region. Then this solution is the only possible solution with these boundary values.

It is easy to prove this uniqueness theorem by considering first of all the potential inside a cavity in a piece of conducting material. The potential on the boundary of the cavity is constant—at a value V0 say. If there is no charge in the cavity, the potential inside it must obey Laplace's equation, $\nabla^2\phi = 0$. One solution of this equation, which satisfies the boundary condition, is that $\phi = V0$ throughout the cavity. This is indeed the only solution. For suppose there were some other solution $\phi 1$ then $\phi 1$ must have at least one maximum or minimum in the cavity, since $\phi 1 = V0$ on the walls, but has a different value somewhere inside. But solutions of Laplace's equation cannot have maxima or minima. At a maximum, for example, $\partial^2\phi/\partial x^2 < 0$, $\partial^2\phi/\partial y^2 < 0$ and $\partial^2\phi/\partial z^2 < 0$, which contradicts the original requirement that



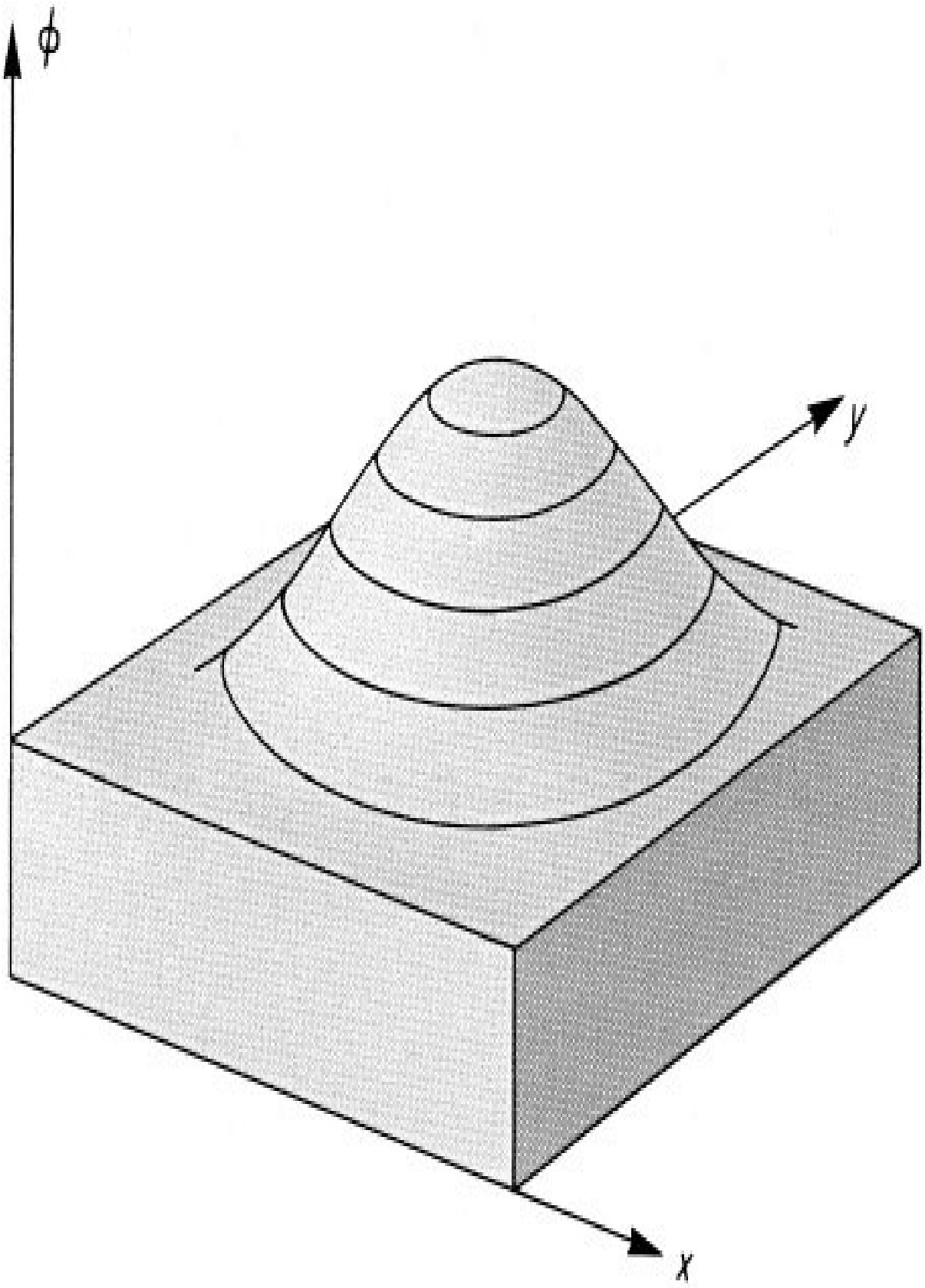
(Think of a hill in a two-dimensional relief map of ϕ, as drawn in Figure 3.2. At the top of the hill both $\partial^2\phi/\partial x^2$ and $\partial^2\phi/\partial y^2$ are less than zero. Similarly at a maximum in three dimensions, $\partial^2\phi/\partial z^2$ is also less than zero.) It follows that there is no other solution than $\phi = V0$ in the cavity.

The proof of the uniqueness theorem can now be extended to show that the theorem holds for Poisson's equation as well as for Laplace's equation, and for any boundary conditions. Imagine that $\phi 1$ and $\phi 2$ are two solutions of Poisson's equation with the same boundary values. The difference $\phi = \phi 1 - \phi 2$ must be zero on the boundary, and must obey Laplace's equation $\nabla^2\phi = 0$ everywhere. One solution to Laplace's equation is $\phi = 0$ everywhere, and we have proved that this must be the only possible solution. It follows that there cannot be different solutions $\phi 1$ and $\phi 2$, and that the uniqueness theorem holds for Poisson's equation also.

Since the potential is constant inside a hollow conductor, the electric field is zero. Sometimes it is useful to put apparatus in a region shielded from electric fields by being enclosed in a conducting box: such a box is called a Faraday cage. The effectiveness of a Faraday cage is a good test of the accuracy of Laplace's equation. Now Laplace's equation was derived from Gauss' law, which was in turn derived from Coulomb's inverse square law. The three statements are in fact equivalent ways of saying the same thing, and a test of Laplace's equation is a roundabout method of testing Coulomb's law. The observation that electrostatic shielding works means that Coulomb's law is certainly valid with a high precision. Changing fields should be shielded as well as static ones, and Coulomb's law can be tested very sensitively by searching for an electric field inside a hollow conductor which is charged up by a high frequency voltage. If Coulomb's law were not quite exact, then a changing field of the same frequency would be found inside the conductor. It is much easier to detect a small high-frequency field than a static one, yet in an experiment at a frequency of a million cycles per second Williams, Faller and Hill* were unable to detect any field within the conductor. If the law of force between charges is not quite inversely proportional to r^2, but varies as l/r^{2+q}, their results show that

q must be less than 10−16.

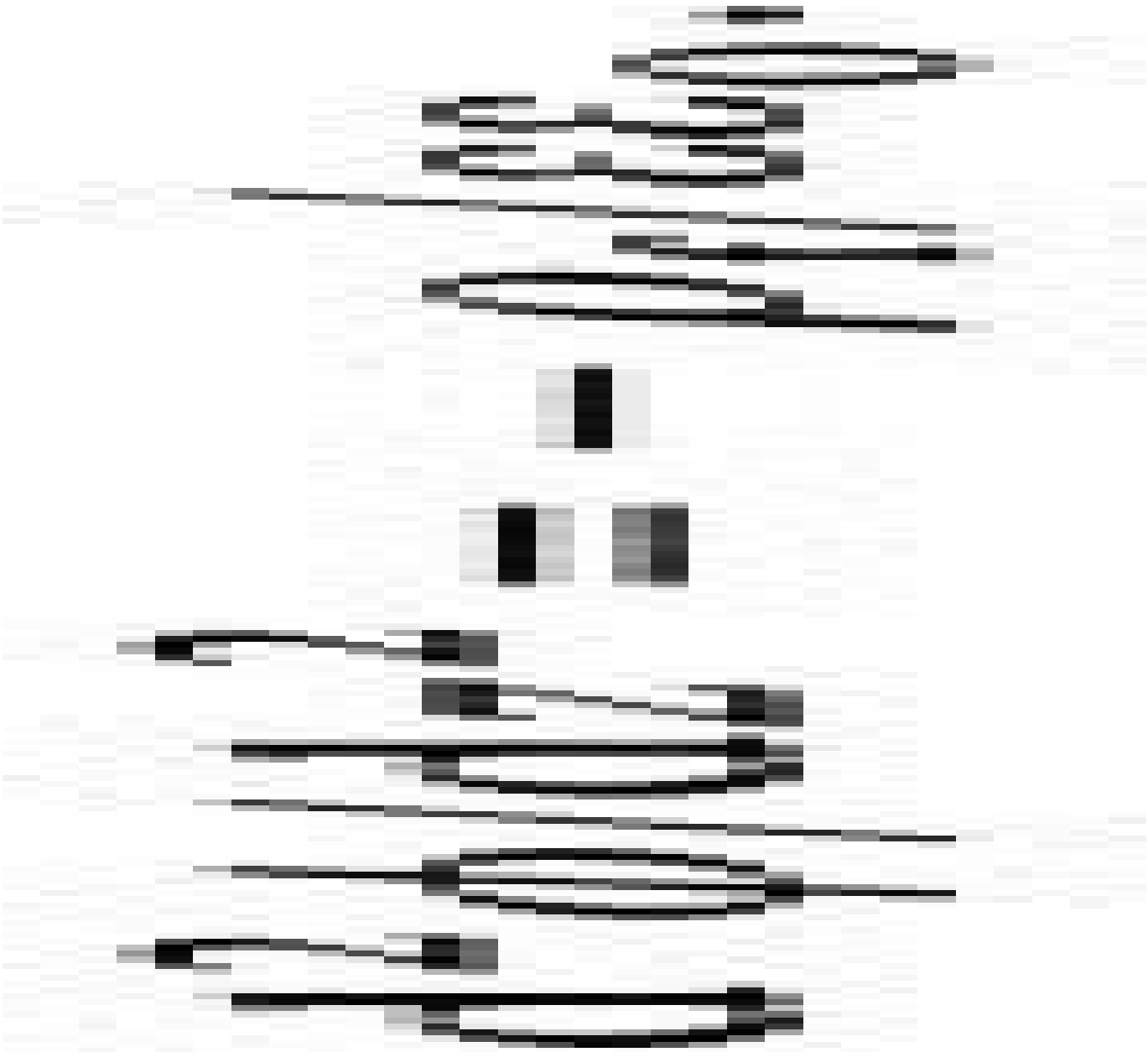
Figure 3.2. A relief map representing a ‘hill’ in a two-dimensional potential.


ϕ
y
x

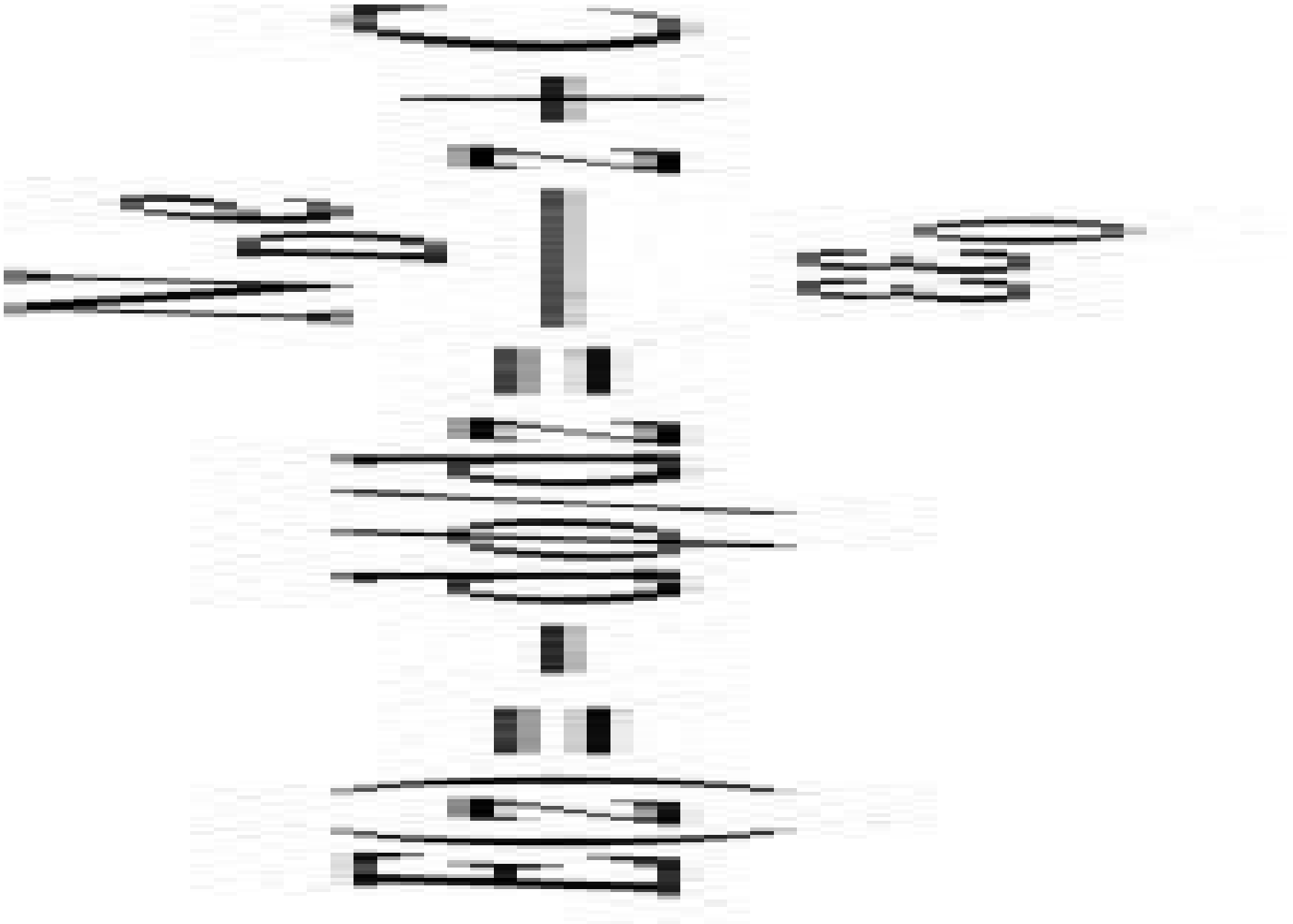
3.1.2 Electric fields in the presence of free charge

An example in which free charge is spread though a macroscopic volume occurs in a metal–semiconductor rectifier. Such a rectifier can be made by coating one face of a silicon disc with gold. The silicon is not pure, but contains a uniform concentration of atoms of a ‘donor’ element such a phosphorus which has a valency of 5, one more than the 4-valent silicon. Each donor atom has an extra electron which is able to move in the silicon and contribute to an electric current. When the gold is maintained at a positive potential with respect to the silicon, a current does indeed flow easily as negatively charged electrons move into the gold. However, when the potential difference is the opposite way round, almost no current flows, i.e. the disc acts as a rectifier, passing current in one direction but not the other. Under reverse bias, electrons are unable to move from the gold into the silicon. As electrons originating from donor atoms close to the metal surface are swept away to the region of high potential, they leave behind a net charge + e for each donor atom. If the density of donor atoms is ND, a layer of charge thus builds up in the silicon with a charge density NDe. Poisson’s equation tells us that the potential is varying within the charged layer, and we must solve Poisson’s equation to find what thickness the layer must be in order to balance the applied potential difference.

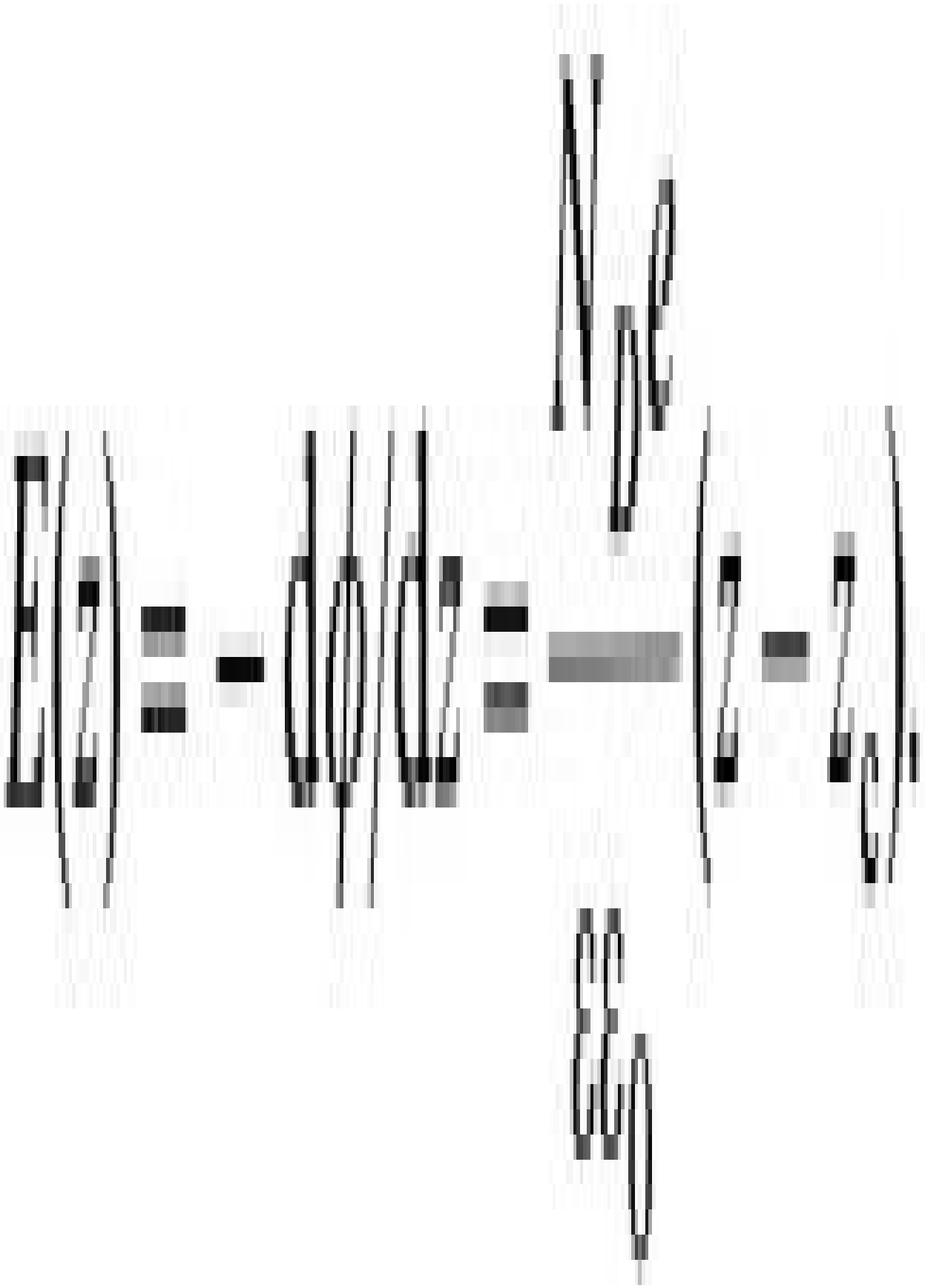
Provided that the thickness of the disc is small compared with its diameter, edge effects can be ignored, and the potential only varies in a direction perpendicular to the disc. Choosing the z-axis to be in this direction, Equation (3.3) simplifies to



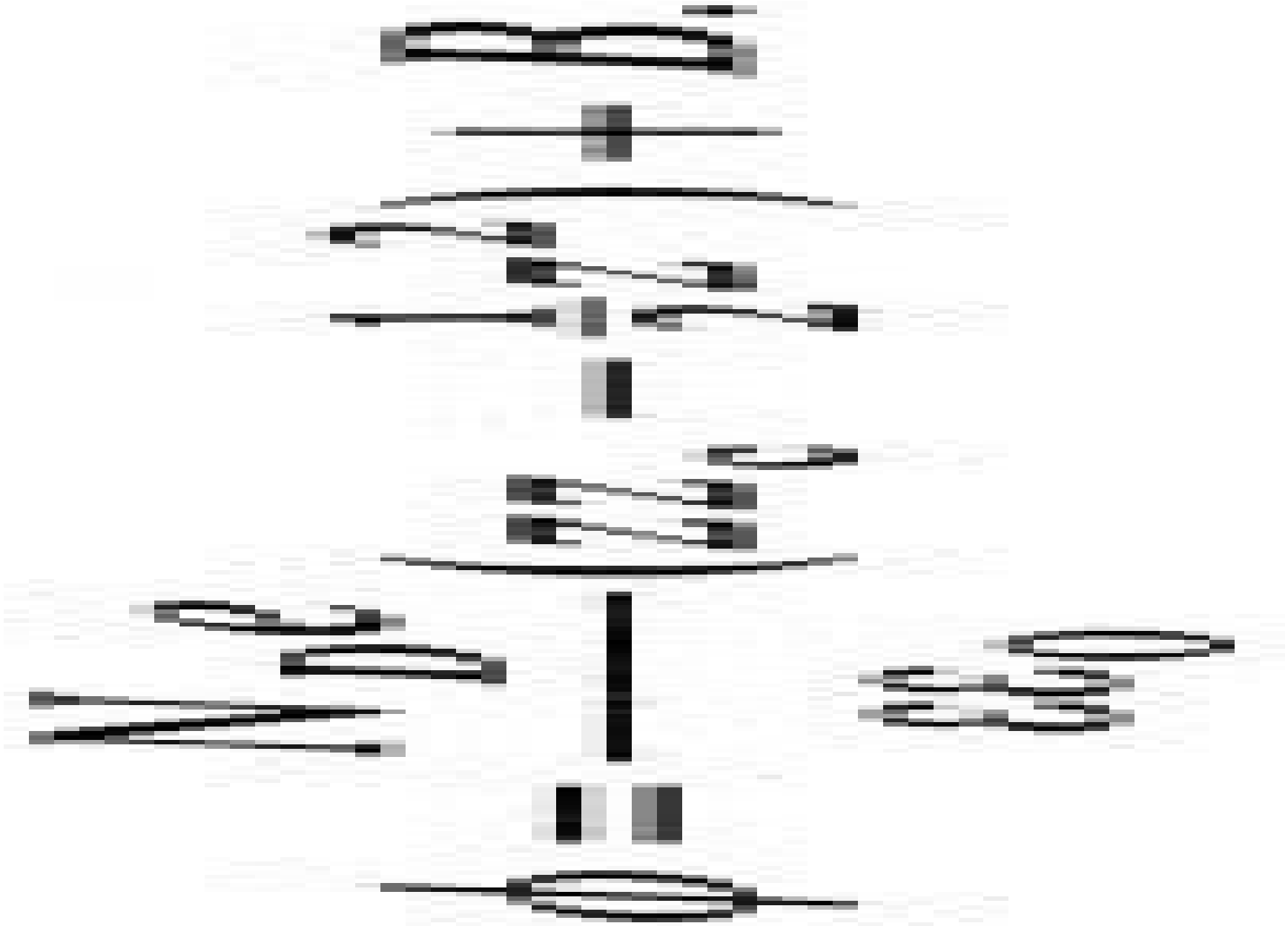
In this example the ‘free’ charge density $\rho_f = ND_e$. The nomenclature is rather confusing here, since the donor atoms are fixed, and the charges they carry are only free in the sense that they do not arise from polarization effects. Polarization charges are indeed present in addition to the free charges, but they are accounted for by the inclusion of the relative permittivity ε in Poisson’s equation. The equation is now easily integrated to give



within the charged layer. Choose z = 0 at the gold surface, and call the thickness of the charged layer zc. We can find the constant C by noting that the field in the silicon falls to zero at the edge of the charged layer, i.e. E(zc) = 0. This condition is satisfied by



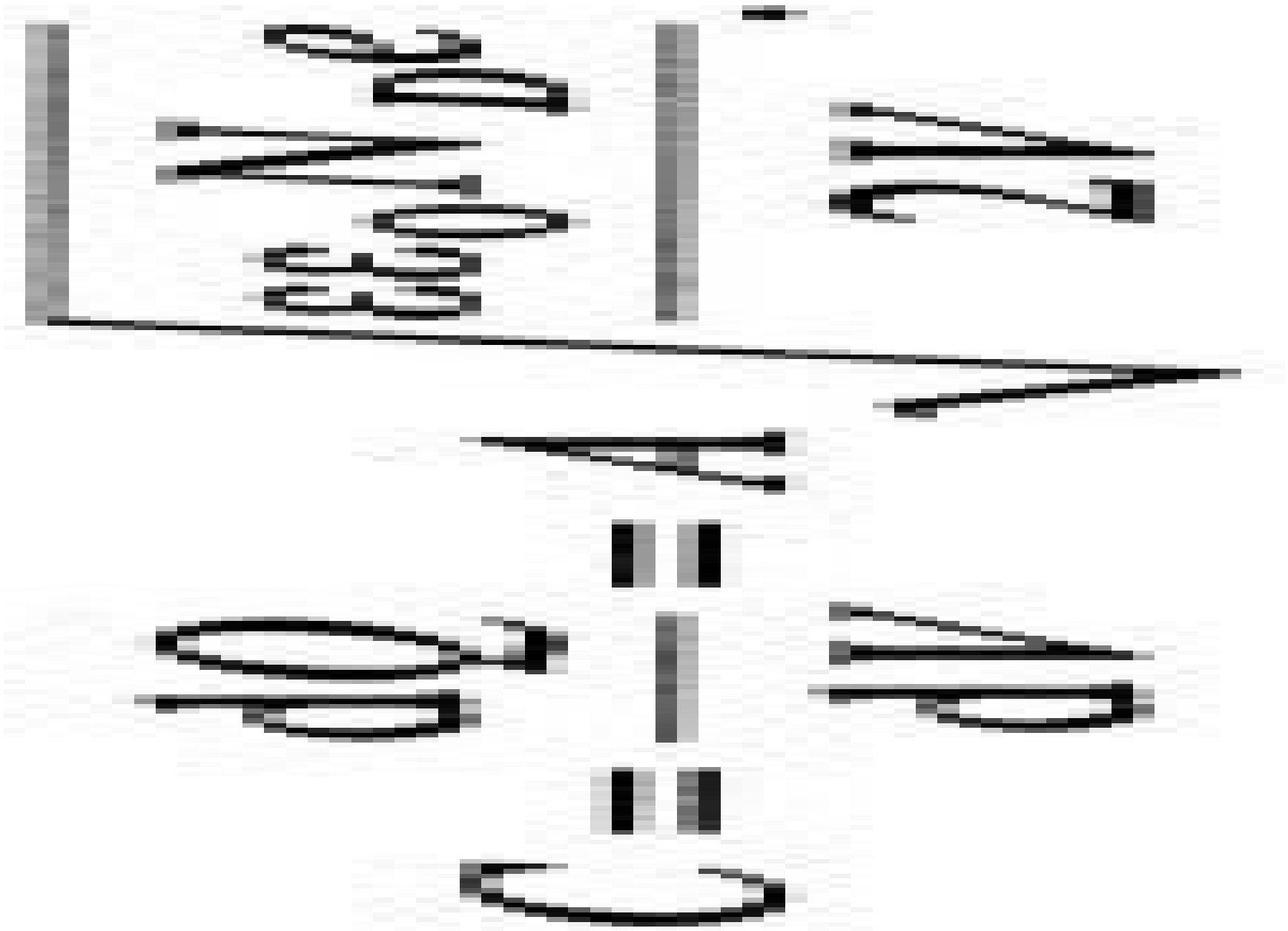
Integrating again,



If the potential is 0 at the gold surface and V at z = zc, B = 0 and .

For silicon, the relative permittivity is 12, and for a specimen with ND = 10^{18} atoms m−3, zc is about mm when V = 100 volts.

If the disc has an area A, the total charge in the layer is Q = NDezcA = . Since there is no electric field outside the disc an equal and opposite charge must accumulate on the gold face. The disc is thus like a capacitor, except that instead of charge and voltage being proportional to one another, the charge increases as the square root of the voltage. The effective capacitance for a small signal is the ratio of the change in charge for a small change in voltage

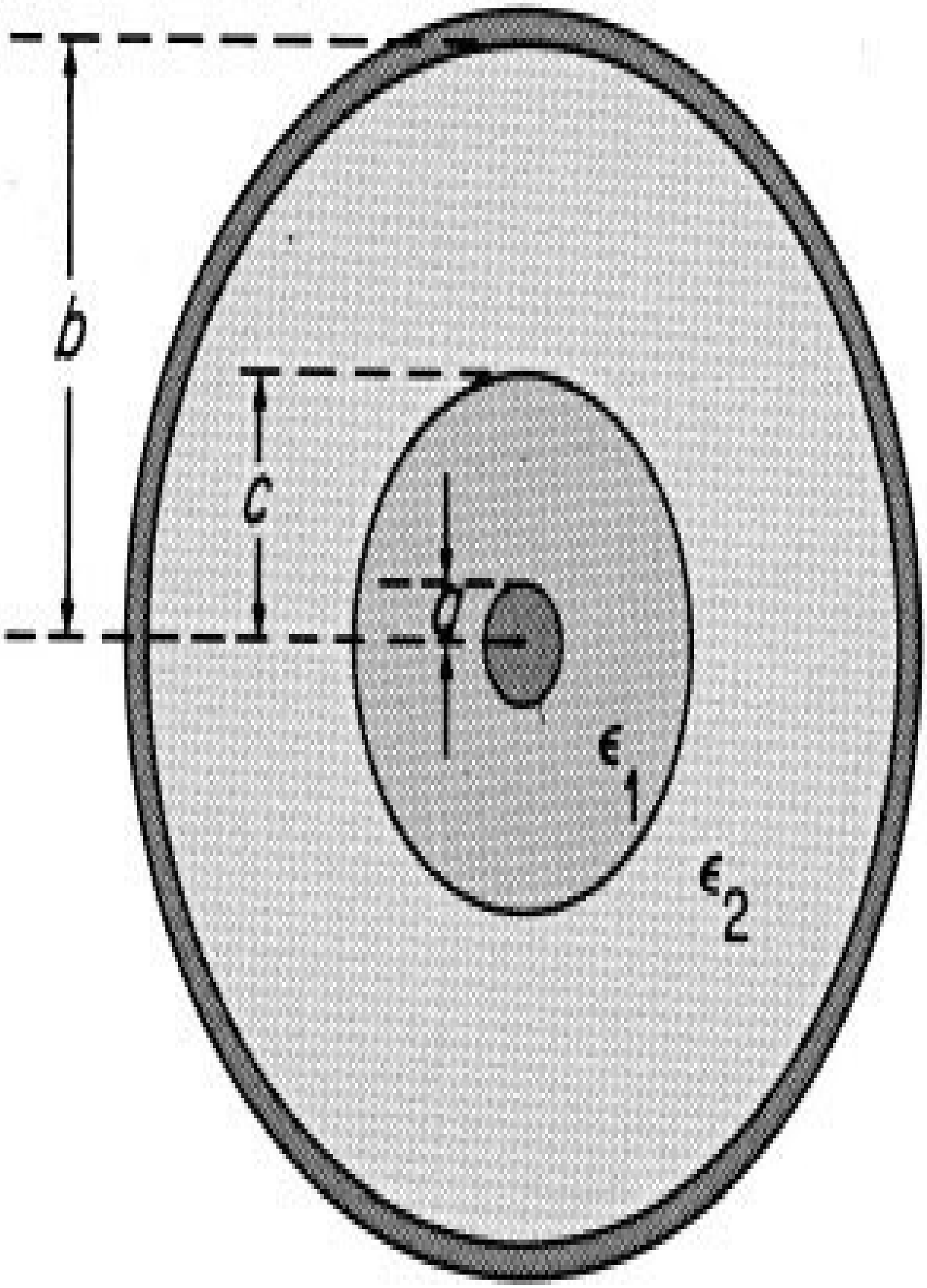


The disc can be used as a variable capacitor with a capacitance controlled electrically by varying the large steady voltage across it. In the example given above, the capacitance for a 1cm² disc is about 30 pF. Variable capacitors operating on this principle are commonly used in high frequency circuits, though they are usually junctions between two differently doped semiconductor layers rather than a metal-semiconductor interface.

3.2 BOUNDARIES BETWEEN DIFFERENT REGIONS

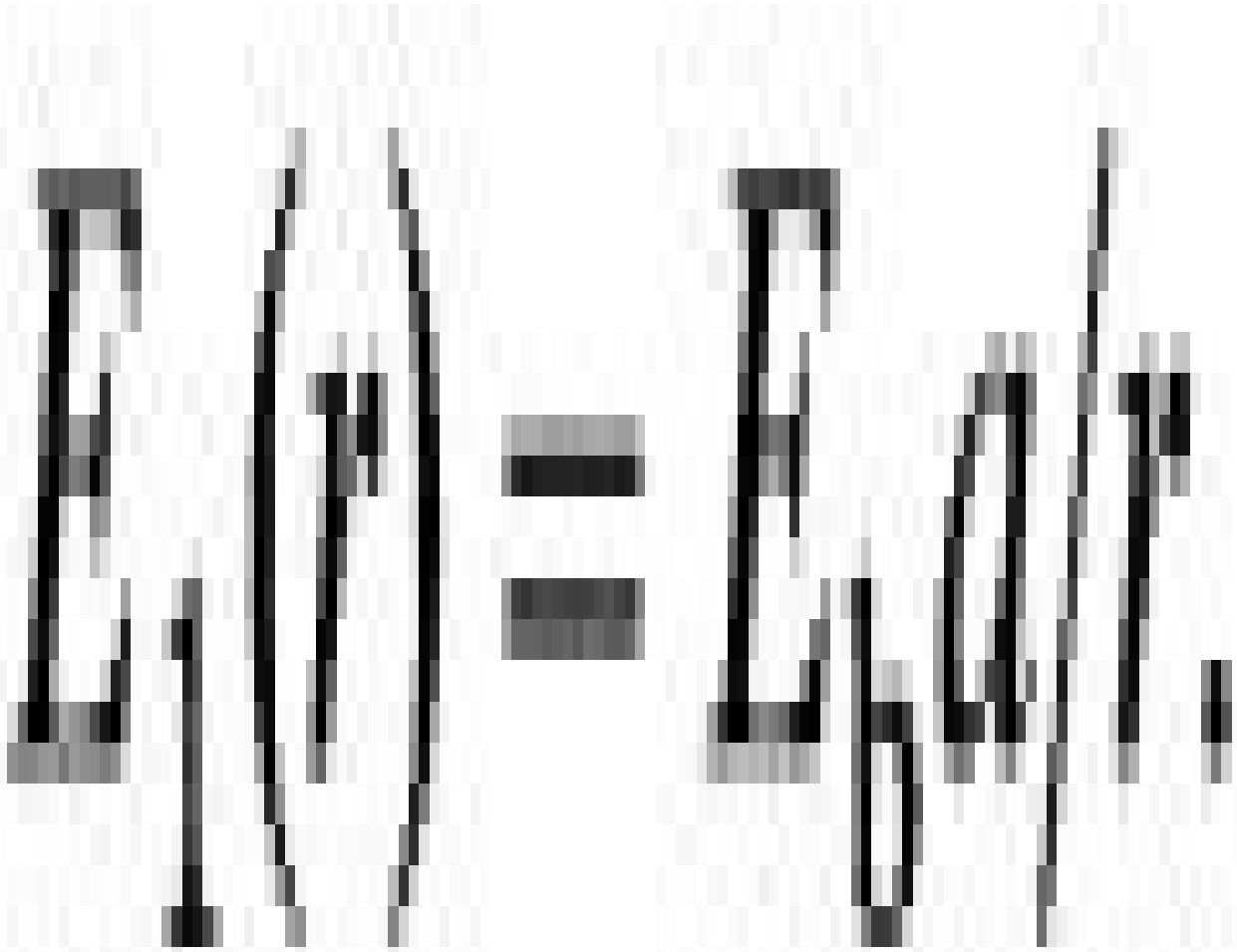
The electric field inside a region of space enclosed by a boundary is generated by charges which may lie inside or outside the region. The field can be derived from a potential which must satisfy Poisson's equation. Because of the uniqueness theorem, the potential inside the region is determined once we know the charge distribution inside the region and the value of the potential on its boundary. For example, the potential must be constant inside a closed conductor containing no charges, although the potential outside may have all kinds of shapes, provided that it is constant on the conducting surface itself. When a situation arises in which there are several regions, each with different properties, we can find the potential in each region separately by solving Poisson's equation or Laplace's equation, subject to a consistent set of boundary conditions. As an example of this procedure, we shall solve Laplace's equation for a coaxial cable containing two materials of different relative permittivity. The insulation between the conductors is made up of two cylindrical sleeves of relative permittivities $\varepsilon 1$ and $\varepsilon 2$, as shown in Figure 3.3. We shall discover that it can be advantageous to make high voltage cables in this way, because if $\varepsilon 1 > \varepsilon 2$, such a cable can sustain a greater potential difference between the conductors without breakdown than a cable of the same size with a single dielectric. Suppose that both dielectric materials break down in an electric field greater than some critical value Eb. As before, the radii of the inner and outer conductors are a and b, and the radius of the interface between the two dielectrics we shall call c.

Figure 3.3. Cross-section of a coaxial cable containing two different dielectric materials.

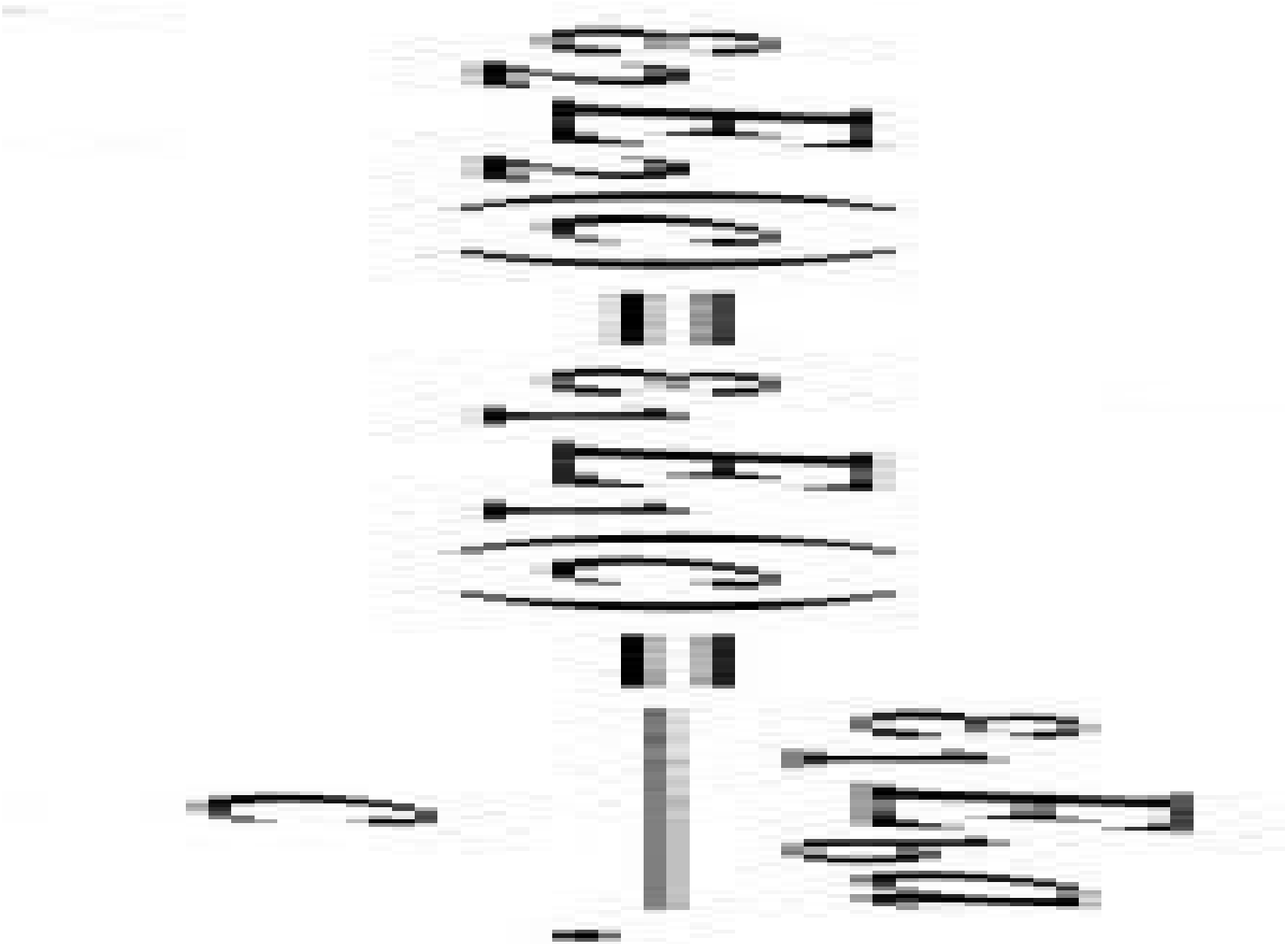

b
c
a
ϵ_1
ϵ_2

What choice of the radius c will allow the cable to carry the maximum possible voltage across its conductors, and what is this voltage?

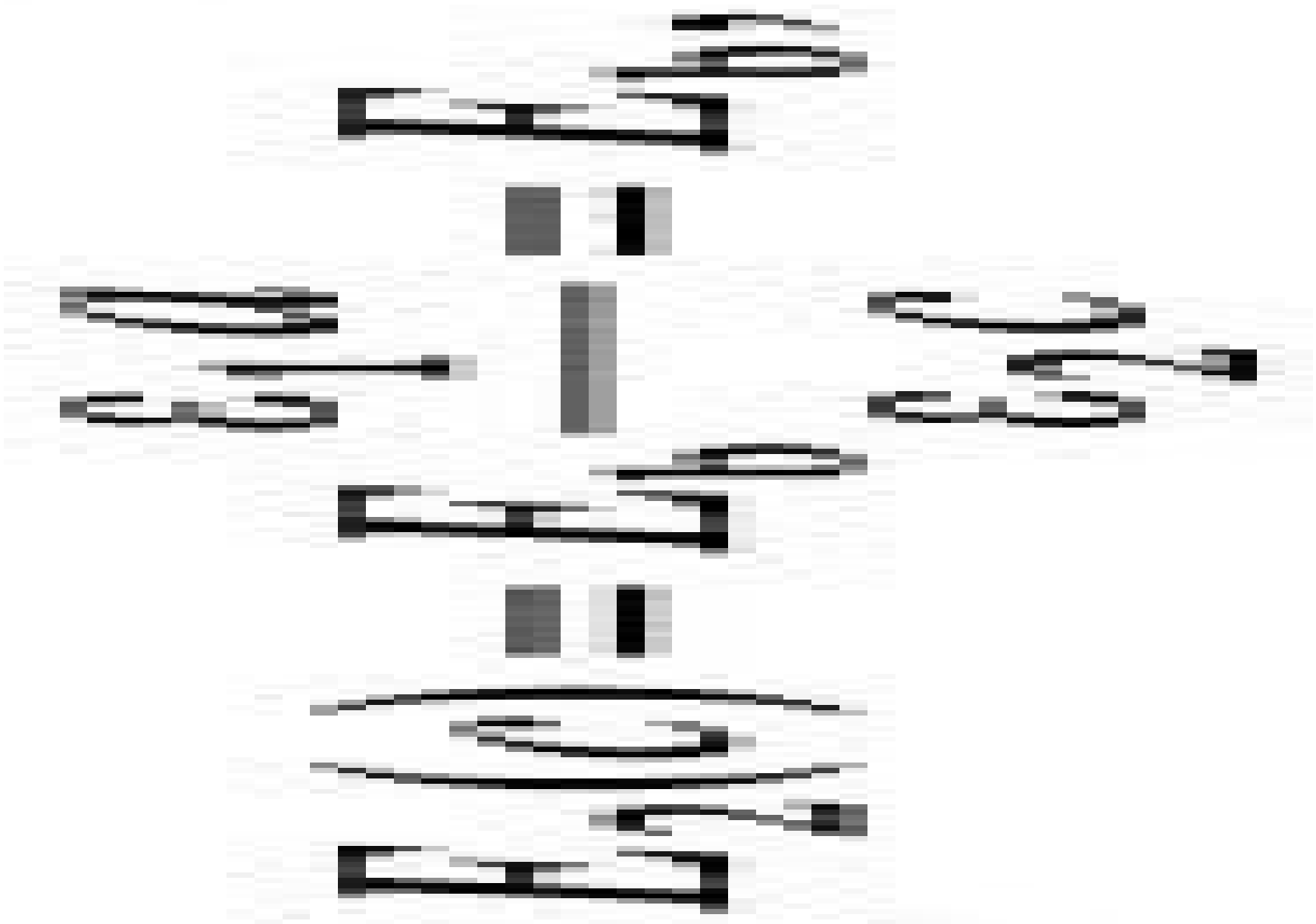
Boundary conditions are not always most conveniently expressed in terms of the potential, and in this example they are in terms of the field, since the condition for breakdown refers to the field. When the inner dielectric is on the point of breakdown, the field must be Eb at the surface of the inner conductor, and the field E1 in the region $a < r < c$ is



We have already shown in section 2.3.3 that at a boundary between dielectrics the perpendicular component of the electric displacement is continuous. In the cable the electric field is perpendicular to the boundary, and the field E2 in the region with relative permittivity ε2 therefore satisfies the condition

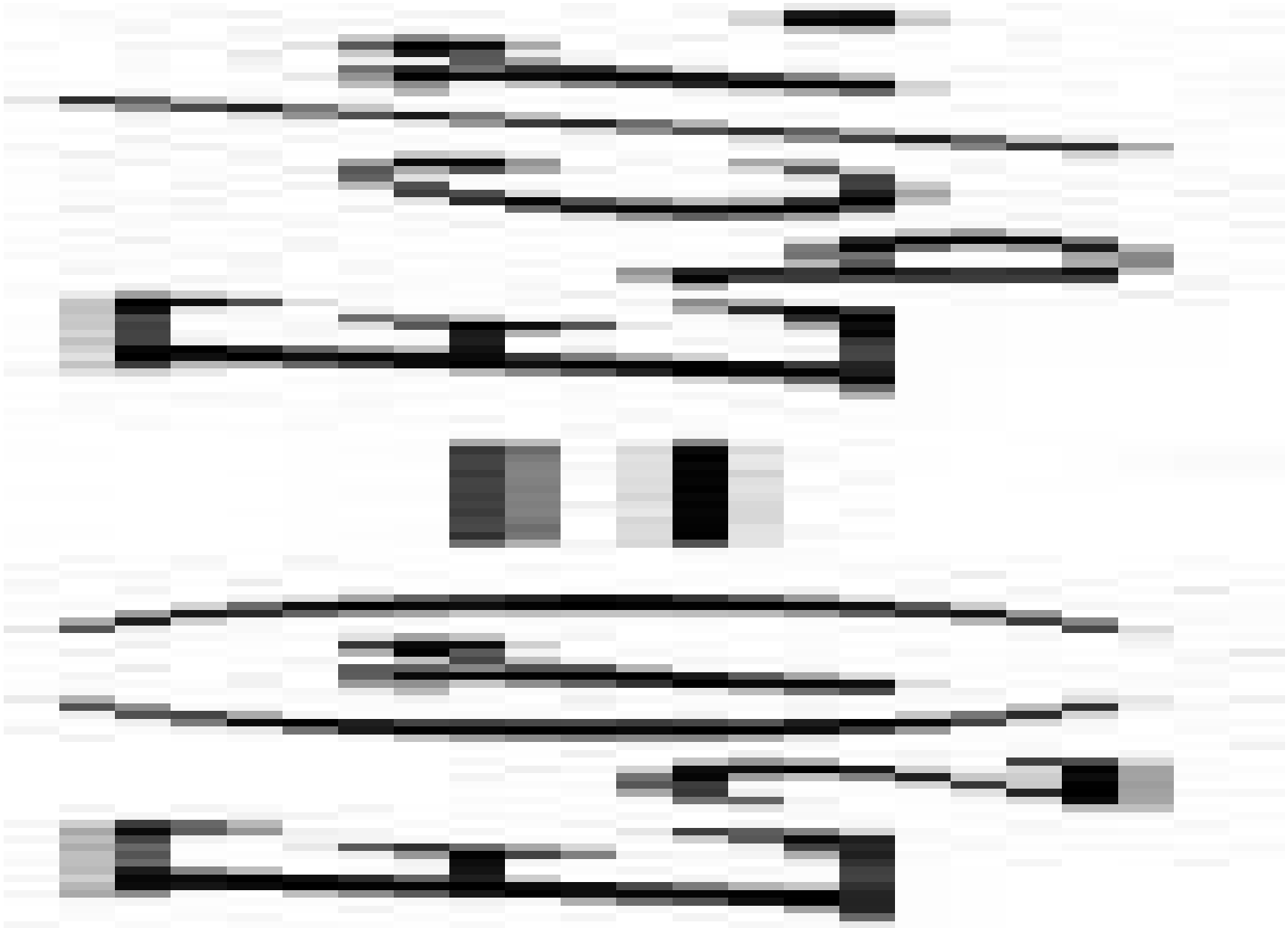


Clearly the highest possible voltage occurs across the cable when both dielectrics are on the point of breakdown, and



or

The field in the outer region is also proportional to l/r, and



Now we know the electric field in the whole of the cable when it is on the point of breakdown, and can evaluate the potential difference between the inner and outer conductors (again choosing the potential to be zero on the outer conductor):

$$V \equiv -\int_b^a \mathbf{E}\cdot d\mathbf{r}$$

$$\equiv -\int_b^c E_b \frac{c}{r} dr - \int_c^a E_b \frac{a}{r} dr$$

$$\equiv E_b \{ c \ln(b/c) + a \ln(c/a) \}.$$

The technique of making high-voltage cables with two dielectrics has occasionally been used for special purposes. Such cables are not in general use, however, because the main practical difficulty in the large scale manufacture of high-voltage cables is in the avoidance of small pockets of air within the dielectric. Since the relative permittivity is low in the air pockets, the electric field is high inside them, and they are sites at which breakdown is likely to be initiated.

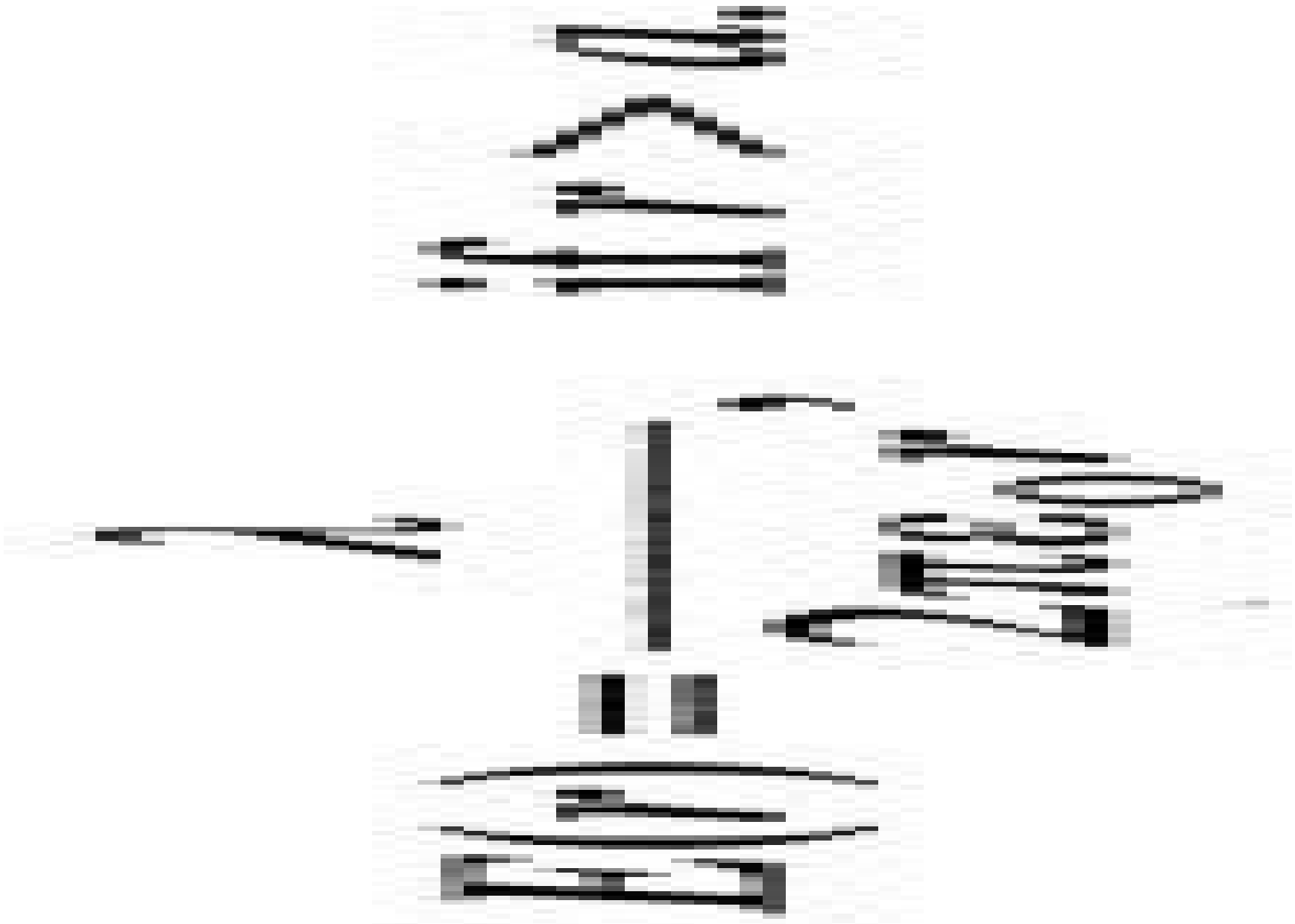
3.3 BOUNDARY CONDITIONS AND FIELD PATTERNS

In the example discussed in the last section of a coaxial cable made up with two dielectrics, the electric field is radial just as it is in a cable containing only a single dielectric. This is of course because the whole cable has axial symmetry, and we can deduce that the field must be radial without knowing anything about boundary conditions at a dielectric interface. Many practical problems are defined by boundary conditions which do not possess a simple symmetry. The electric field and potential must then adopt a pattern which satisfies these boundary conditions. In this section we discuss some examples in which the direction of the field lines cannot be deduced using symmetry, but which do have exact analytical solutions.

3.3.1 Electrostatic images

When a conductor is placed in an electric field, induced charges appear on its surface, distributed in such a way that the whole of the conducting region is at the same potential. However, we do not need to know what goes on inside the conductor. Provided that the conducting surface remains an equipotential, the uniqueness theorem tells us that there is only one possible solution for the field outside. Sometimes calculations can be simplified by imagining that the conductor is not present. The actual induced charges can now be redistributed inside, so long as the potential at the surface is not changed. Let us first illustrate this with an example in which the field does have a simple symmetry. If a long conducting wire is charged up, and is far away from other charges or conductors, the field around it is axially symmetrical. The magnitude E(r) of the field at a distance r from the axis of the wire can therefore be found by applying the integral form of Gauss' law to a cylindrical surface of radius r. If the radius of the wire is a, and it carries a charge λ per unit length, then the flux of E out of unit length of the cylindrical surface is

and thus

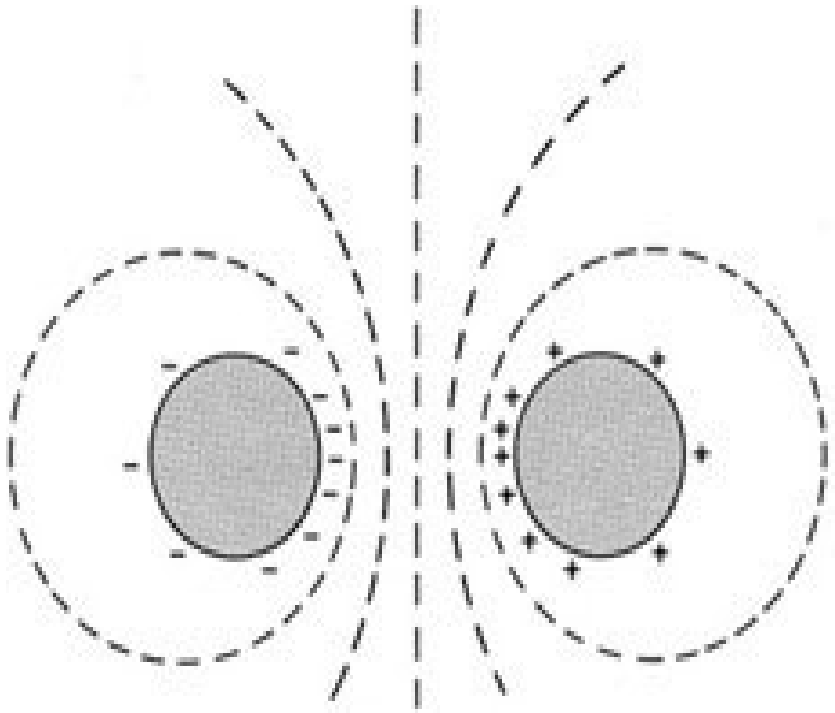


Integrating this equation we find that the potential is

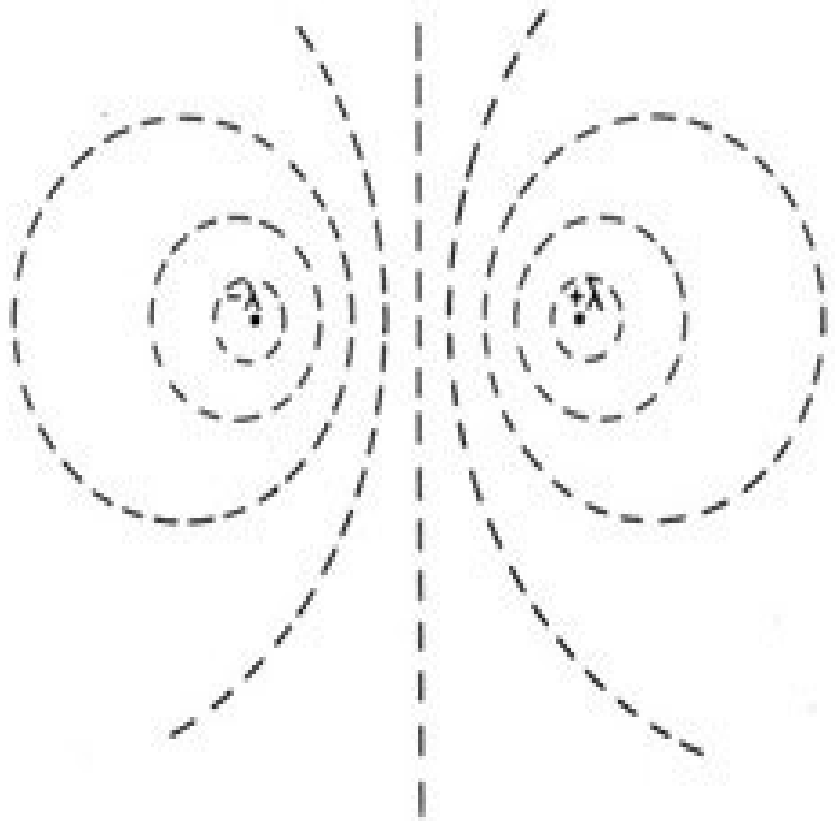
$$\phi(r) = \frac{-\lambda}{2\pi\varepsilon_0} \ln(r) + \text{constant}$$

$$= \frac{\lambda}{2\pi\varepsilon_0} \ln\left(\frac{b}{r}\right),$$

Figure 3.4. Equipotentials around a pair of parallel cylindrical conductors, and the equivalent image system.

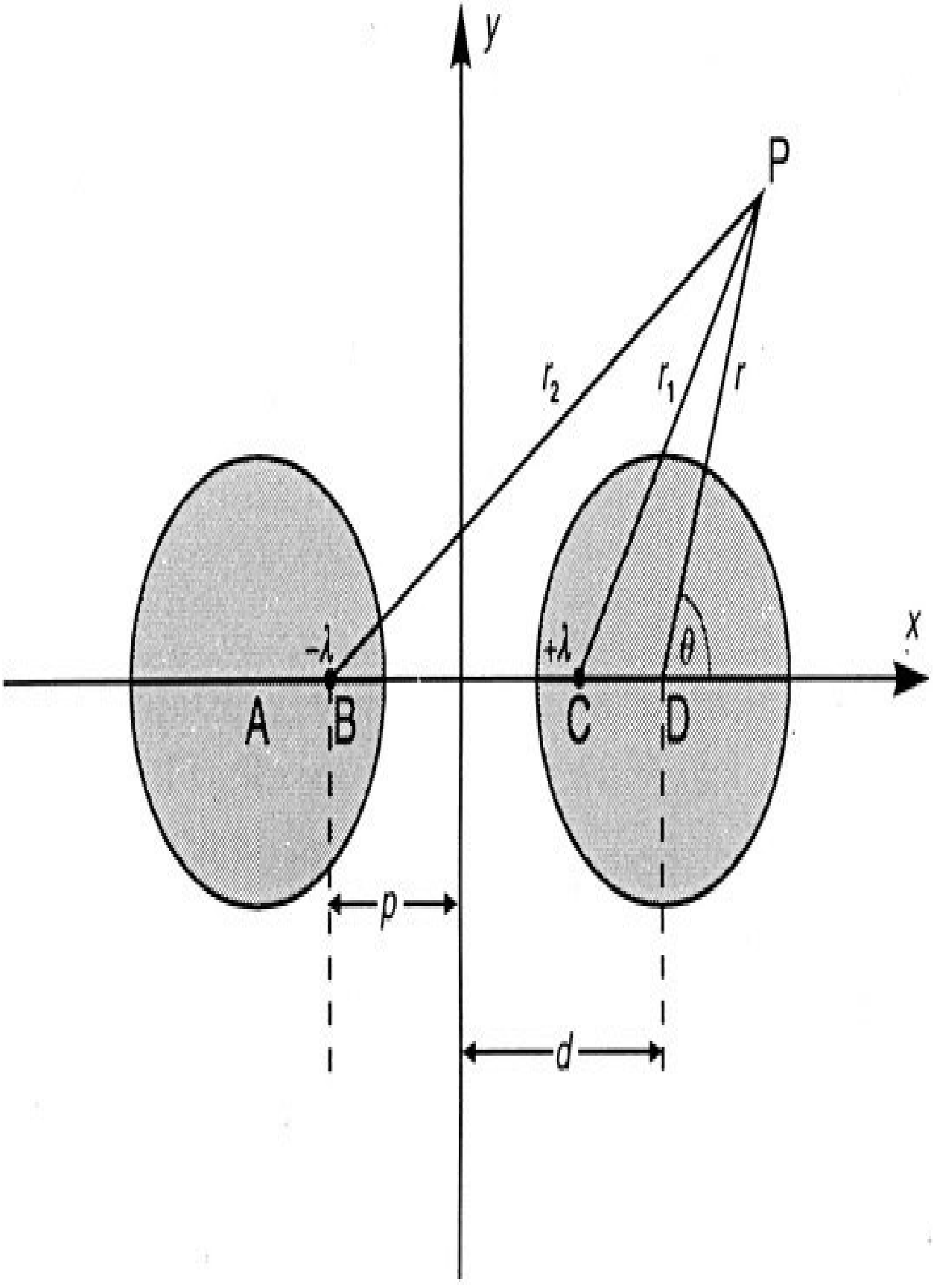


(*a*) Charge is distributed non-uniformly on the surfaces of the conductors.



(*b*) The image charges are collapsed onto lines.

Figure 3.5. Working out the potential due to the image charges.


y
P
r_2
r_1
r
$-\lambda$
$+\lambda$
θ
x
A
B
C
D
p
d

choosing the potential to be zero* at r = b. This expression is only valid for r > a, since inside the conductor the potential has the constant value ($\lambda/2\pi\varepsilon 0$) In (b/a). But outside r = a the potential is independent of a, and would be the same if the conductor were collapsed onto the axis to form a line charge λ/unit length. The fictitious line charge is called the ‘image’ of the real surface charges on the conductor.

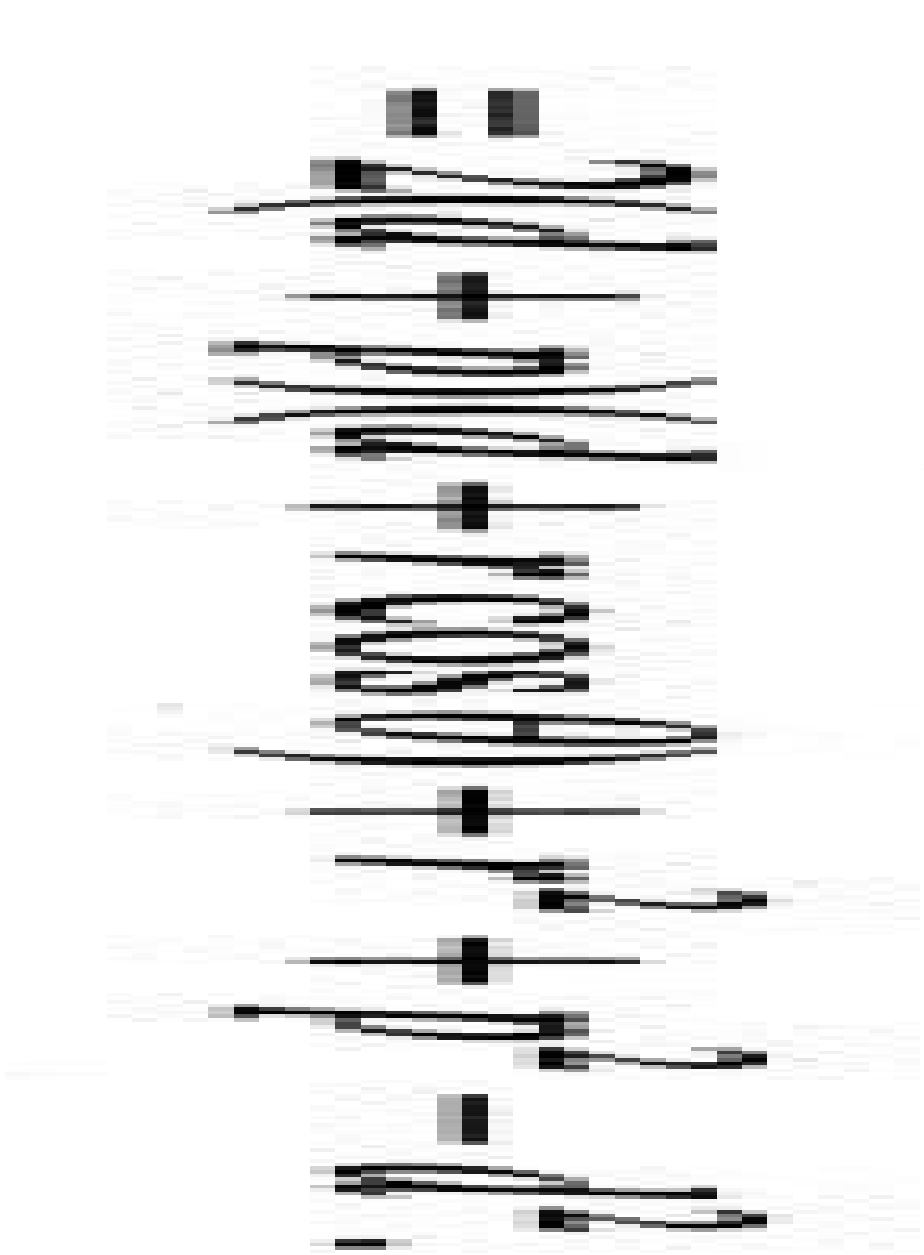
For a single conductor, we do not learn anything new by introducing the image charge. The image method is helpful, however, in finding the field around two parallel cylindrical conductors carrying equal and opposite charges. We shall prove shortly that the equipotential surfaces around the conductors are cylindrical, as drawn in Figure 3.4(a) The equipotentials are not coaxial with the conductors; the electric field, and hence also the charge density, is not uniform on the conducting surface. Nevertheless the conductors can be represented by the image system shown in Figure 3.4(b), in which the surface charge is collapsed onto two lines carrying charges ±λ per unit length. Let us calculate the potential generated by these images. A section through the parallel pair of conductors is shown in Figure 3.5.

The conductors are each of radius a, and their centres at A and D on the x-axis are at distance d from the origin. Image charges ±λ per unit length lie along the conductors, crossing the x-axis at C and B respectively. The distances ±p of the line charges from the origin can be chosen so that the surfaces of the conductors lie on equipotentials of the image charges. Choosing the potential to be zero at infinity* and on the y-axis, the potential at the point P in the x–y plane is

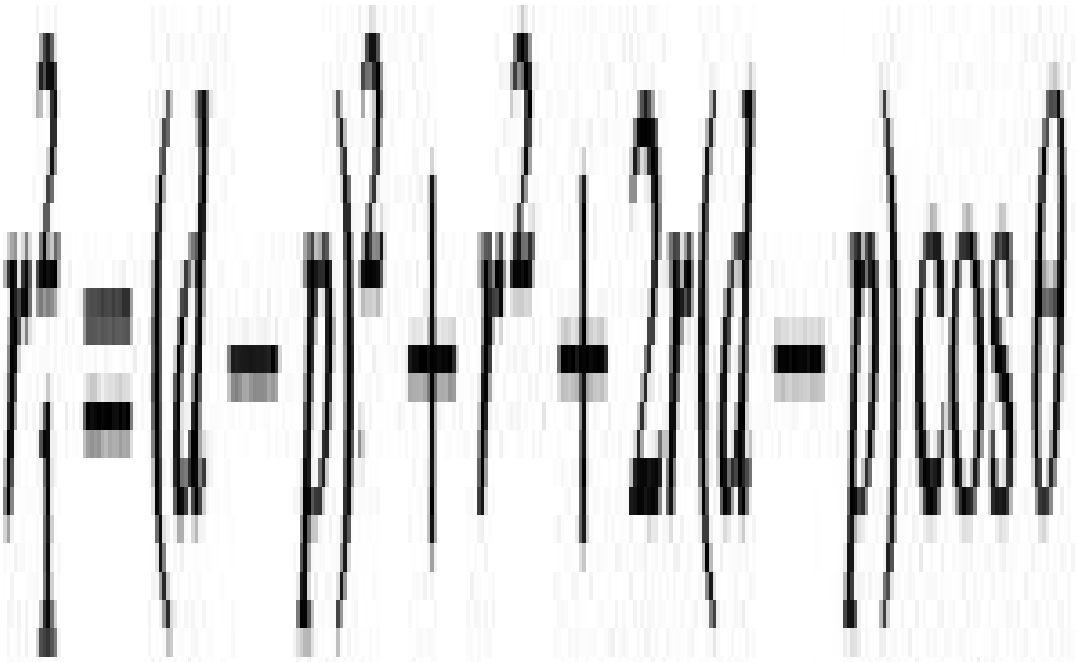
$$\phi_P = \frac{\lambda}{2\pi\varepsilon_0}\ln(1/r_1) - \frac{\lambda}{2\pi\varepsilon_0}\ln(1/r_2)$$

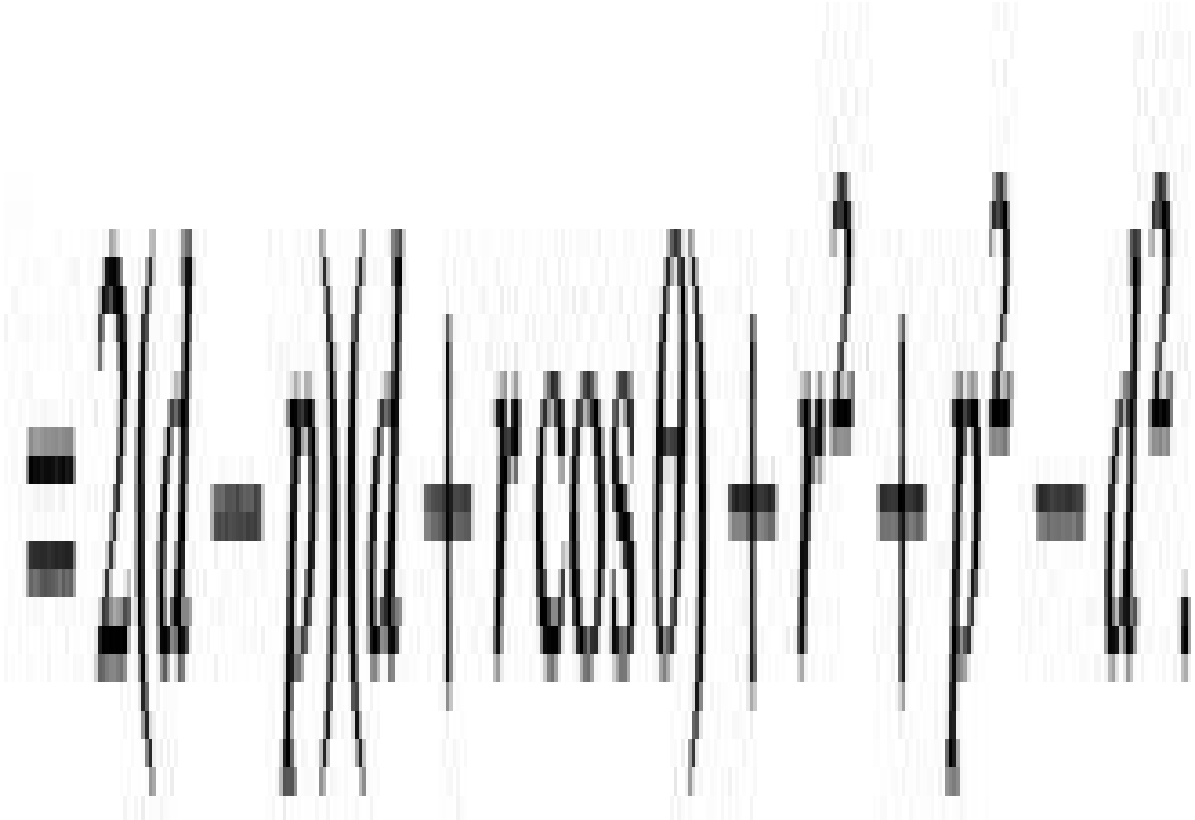
$$= \frac{\lambda}{4\pi\varepsilon_0}\ln\left(\frac{r_2^2}{r_1^2}\right).$$

Now express this potential in terms of the distance r from P to the axis of the conductor D and the angle θ between PD and the x-axis. In the triangle BPD,

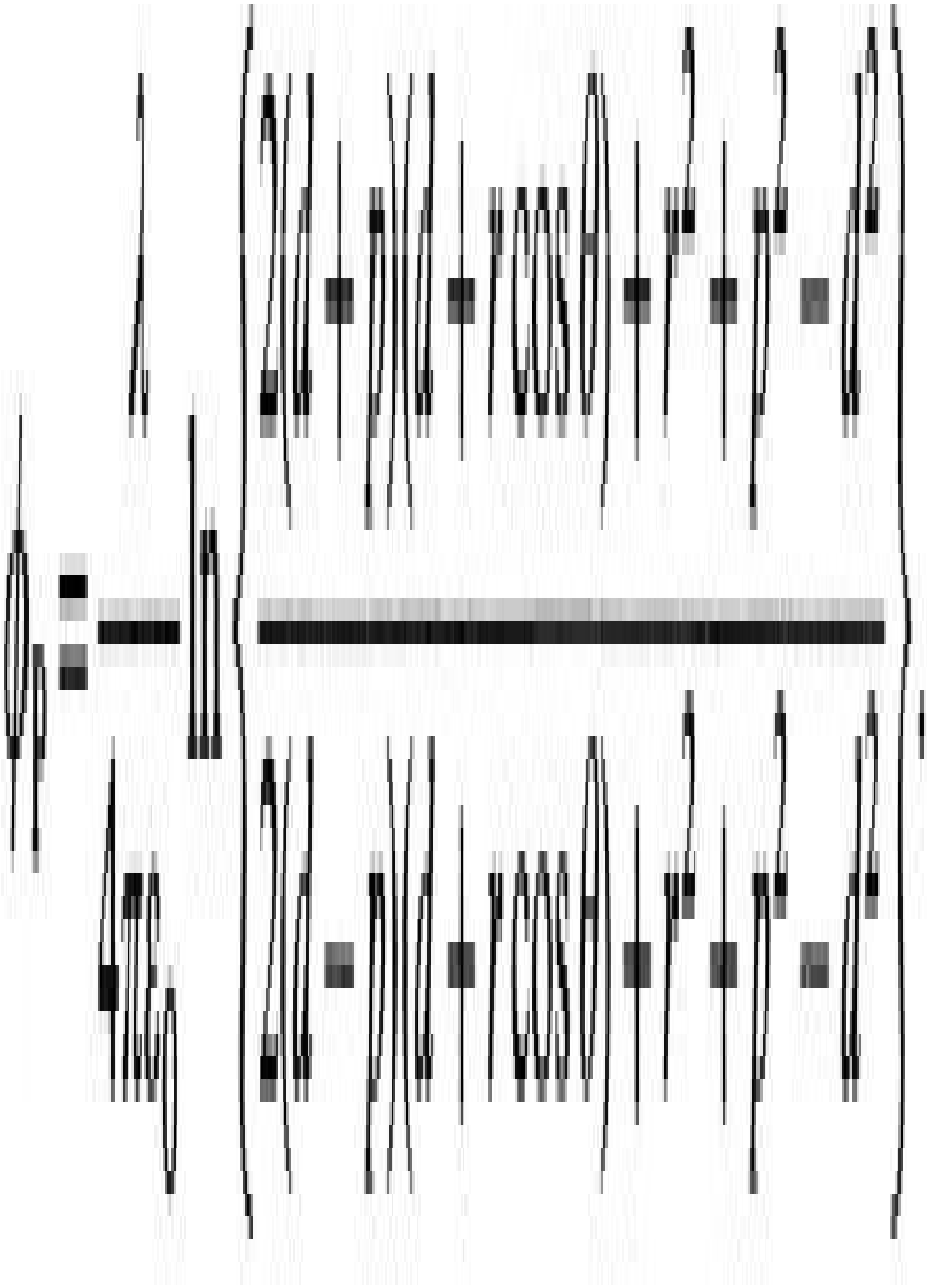


Similarly in triangle CPD,

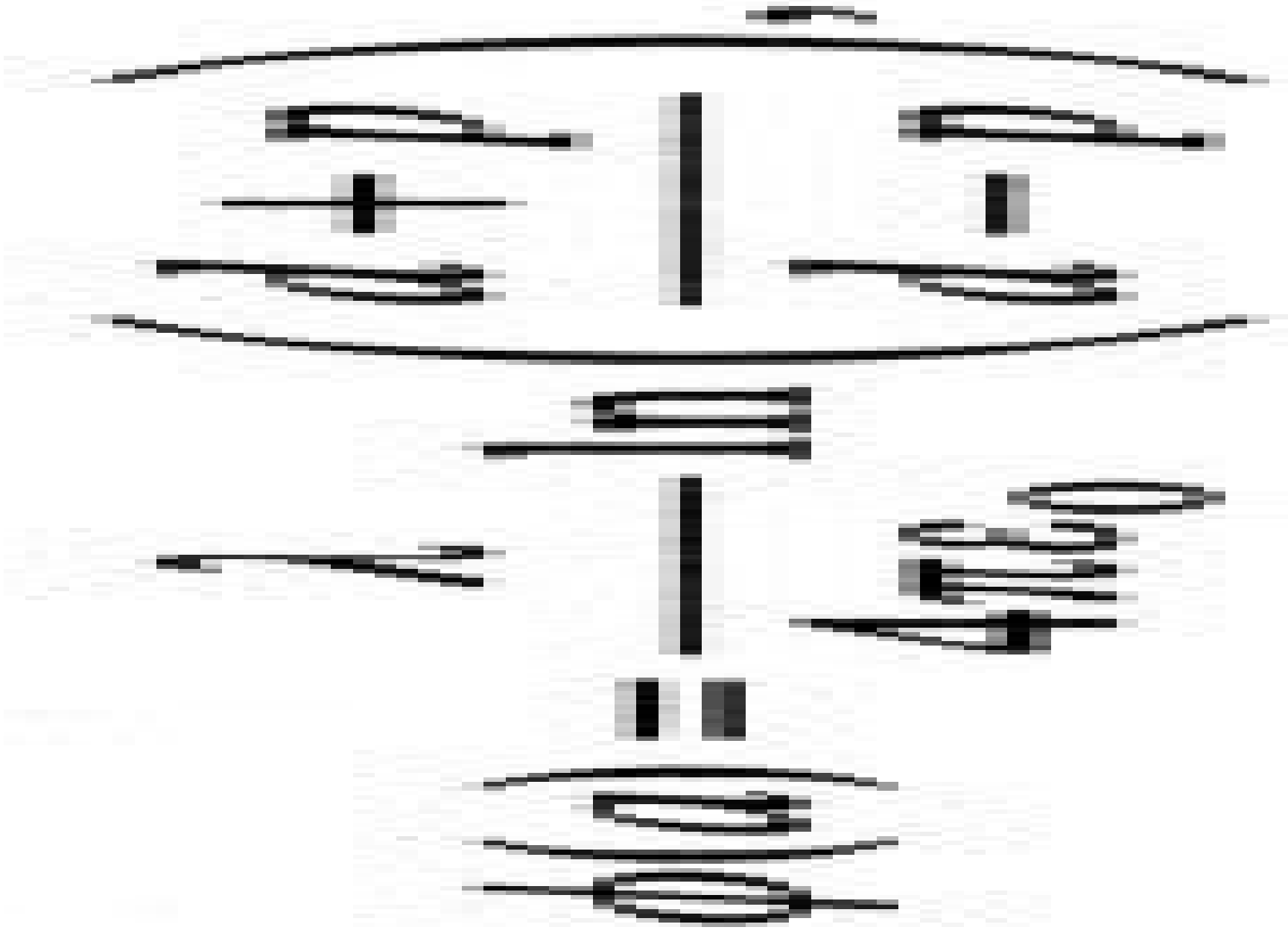




Hence

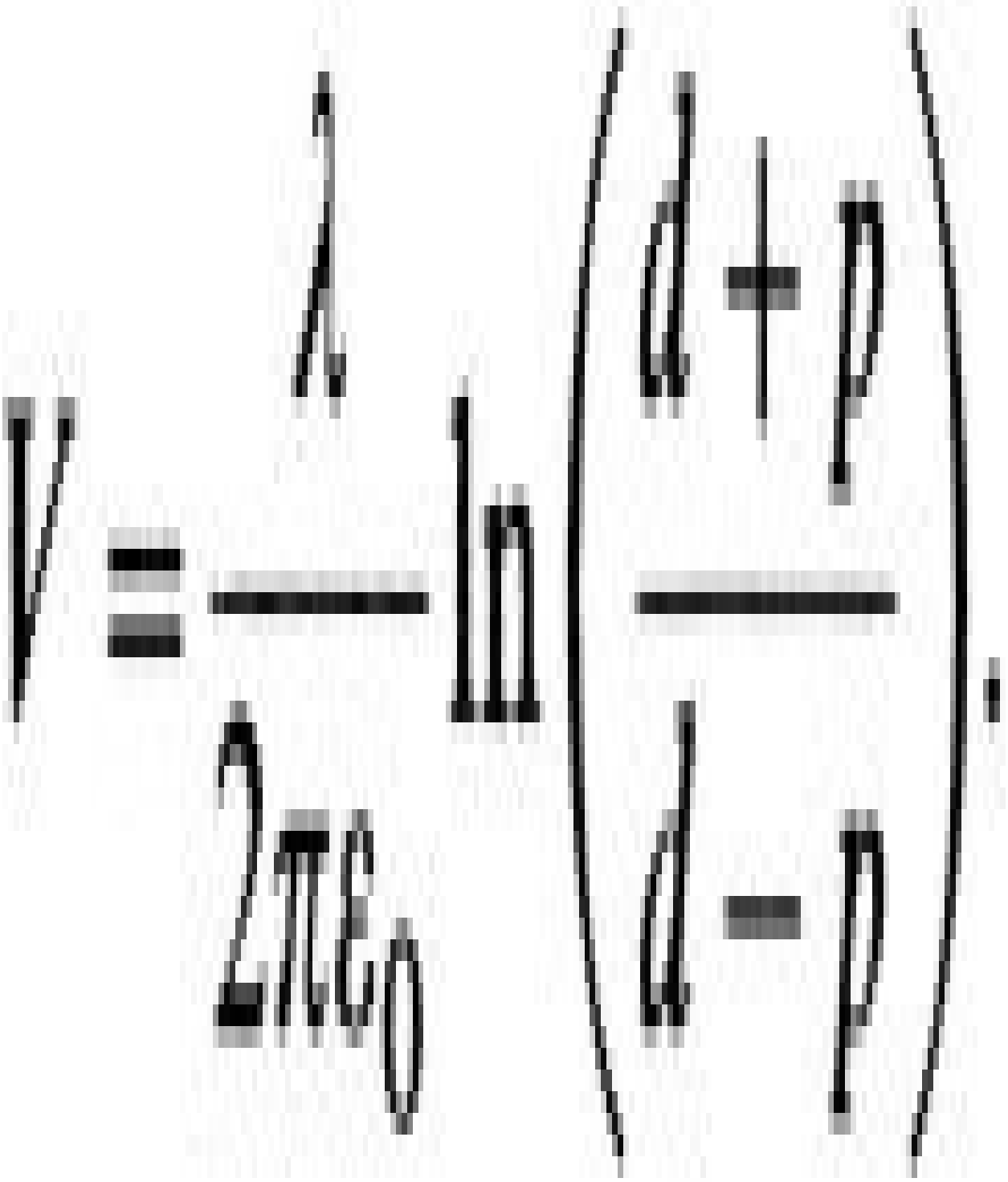


If the position of the image charge is such that $p^2 = (d^2 – a^2)$, then at r = a



independent of θ, and the surface of the conductor with axis through D is an equipotential of the image charge.

By symmetry, the potential at the surface of the other conductor is $-\phi(a)$, and the potential difference between the conductors is



The uniqueness theorem tells us that the potential distribution between the conductors is the same as for the image charges. The total magnitude of the surface charge on each conductor is indeed equal to the image charge, as can be seen by applying Gauss' theorem to one of the equipotentials surrounding each conductor. Knowing the charge carried by the conductors and potential difference between them, we can now calculate their capacitance per unit length, which is

$$C = \frac{\lambda}{V} = \frac{2\pi\varepsilon_0}{\ln\left(\frac{d+p}{d-p}\right)} = \frac{2\pi\varepsilon_0}{\ln\left[\frac{(d+p)^2}{d^2-p^2}\right]} = \frac{\pi\varepsilon_0}{\ln\left(\frac{d+p}{a}\right)}$$

$$\simeq \frac{\pi\varepsilon_0}{\ln\left(\frac{2d}{a}\right)} \quad \text{if } d \gg a.$$

For example, if 2d/a is 6, the capacitance between the pair of wires is about 15pF/m. We shall need to know this capacitance in Chapter 9 when discussing the transmission of electrical signals. It is of the same order of magnitude as the capacitance per unit length of a typical coaxial cable, which we worked out in section 2.3.2.

The method of electrostatic images has here led us to a neat solution of a complex problem, but as a technique for solving Poisson's equation outside conductors, it only works backwards. If we are lucky enough to find a charge distribution which has equipotentials to fit the conducting surfaces, then this distribution can be used as an image of the real surface charges. There is however no systematic way of finding a distribution of image charges equivalent to an arbitrary system of conductors.

3.3.2 Spheres and spherical cavities in a uniform external field

At the surface of a conductor, the potential meets the straightforward requirement that it must be constant. The boundary conditions at the surface of a dielectric are more complicated, and potential problems are correspondingly more difficult to solve when dielectric materials are present. One of the few examples for which the field has a simple form is a dielectric sphere placed in a uniform external field Eext.

We shall show that the dielectric is uniformly polarized, and that outside the dielectric the original field Eext is modified by a field which can be represented by a point dipole at the centre of the sphere. The uniform polarization is a special property of the spherical shape of the dielectric: in general, dielectric material is not uniformly polarized when placed in a uniform external field.

For the spherical dielectric, the field inside has a magnitude Eint, different from Eext. Taking the potential to be zero at the centre of the sphere, and choosing the z-axis in the direction of the field as in Figure 3.6, the potential inside the sphere at a point with spherical polar coordinates (r, θ) is

(3.6)

If the susceptibility of the dielectric is χE, it has a dipole moment per unit volume P = χEε0Eint, and the total dipole moment of the whole sphere is

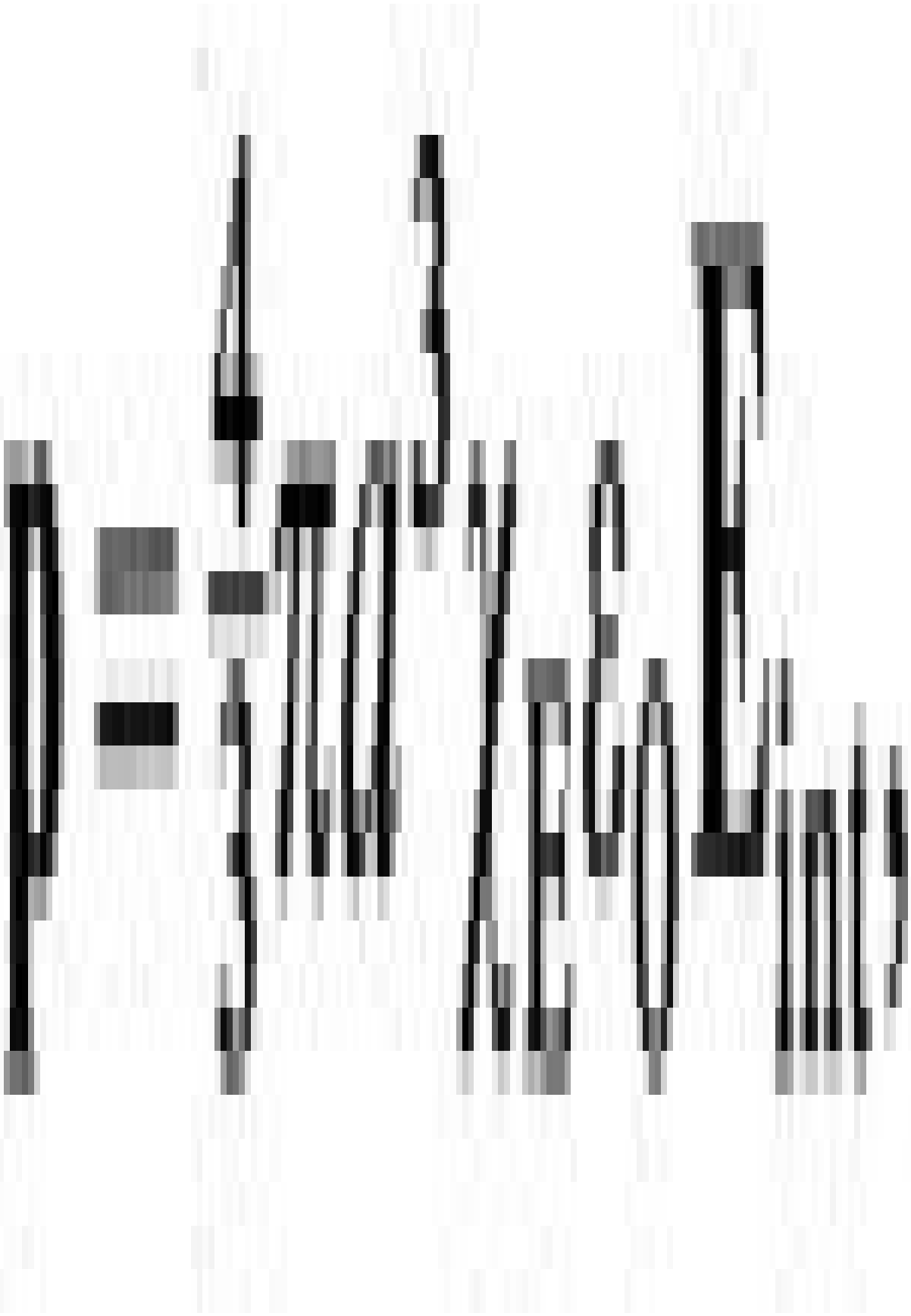
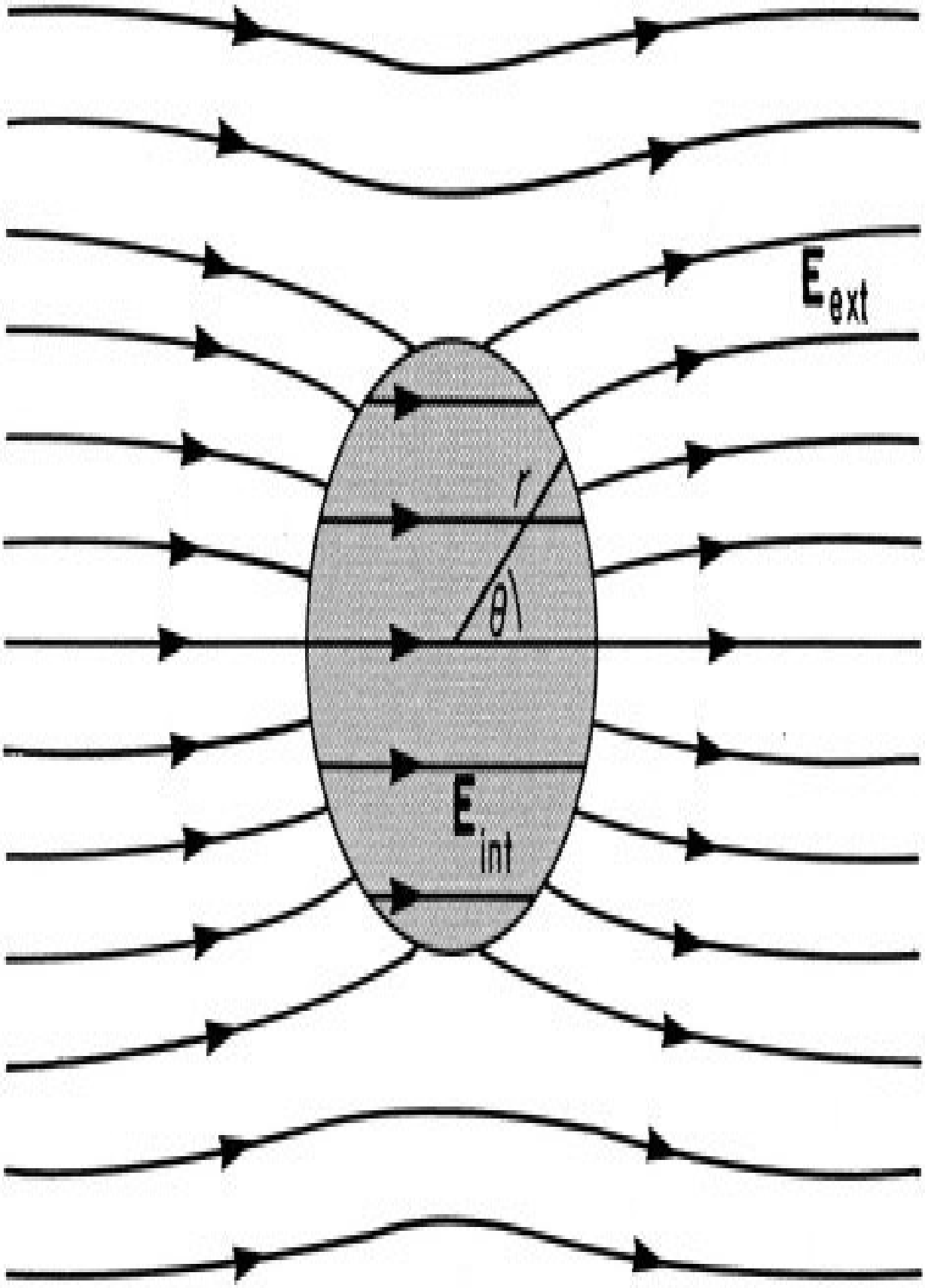
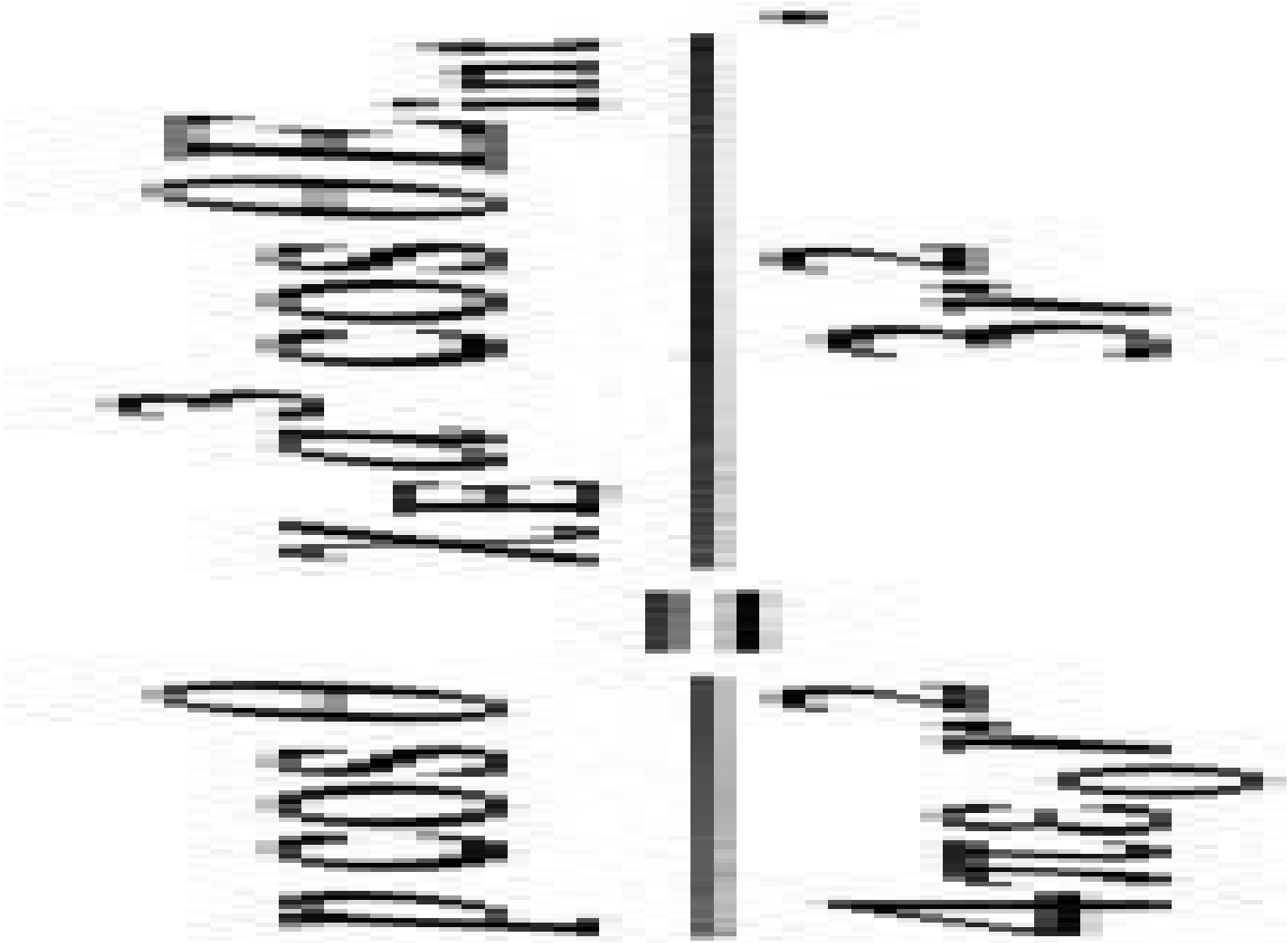


Figure 3.6. A uniform external field is distorted by a dielectric sphere.


E_{ext}
r
θ
E_{int}

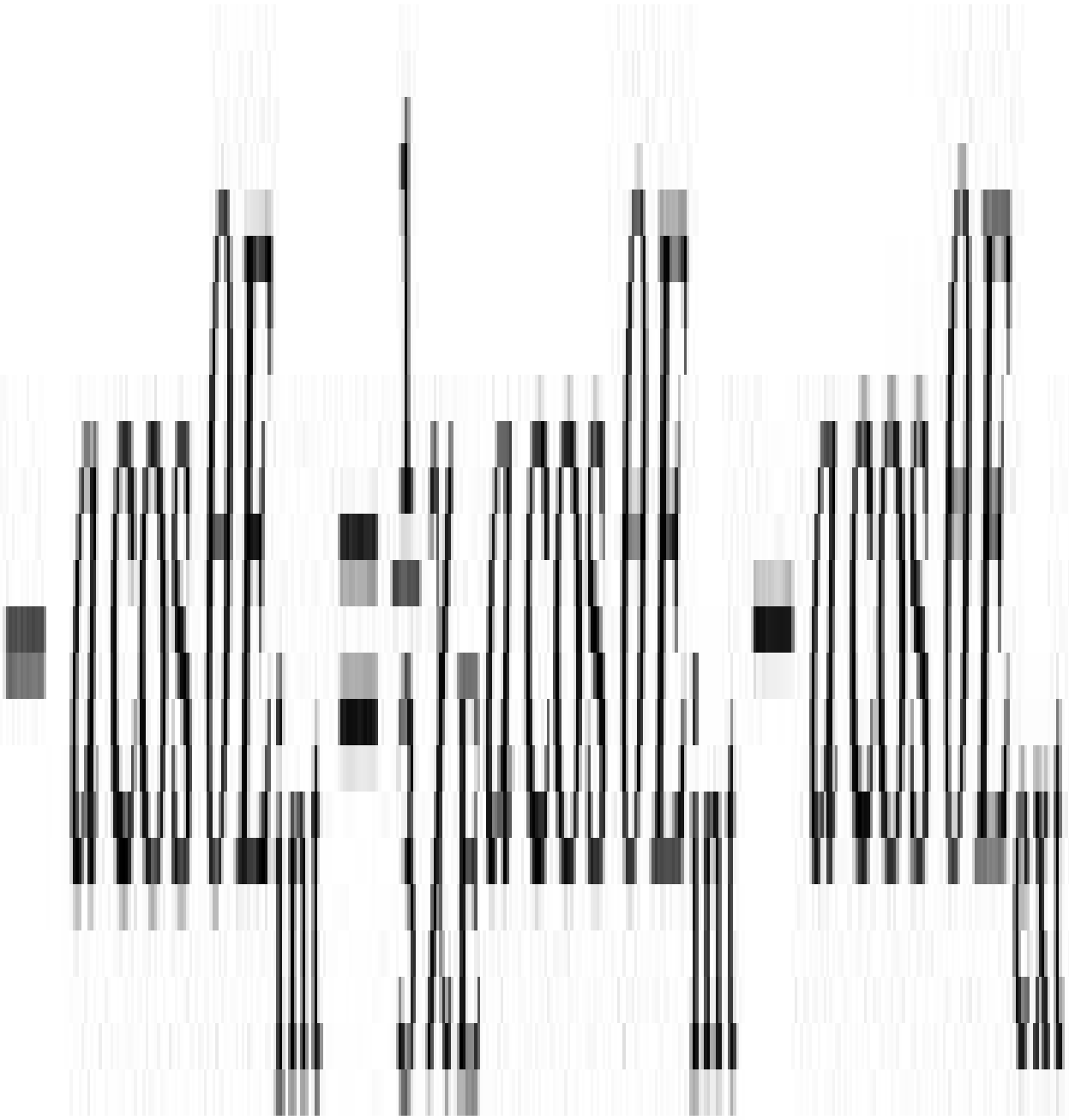
where a is the radius of the sphere. In Section 1.5.3 it was shown that the potential due to a point dipole with dipole moment p pointing in the z-direction is



The potential outside the sphere is the sum of that due to the external field and the dipole potential, i.e.

(3.7)

At the surface of the sphere, where r = a, this potential matches the internal potential provided that



or

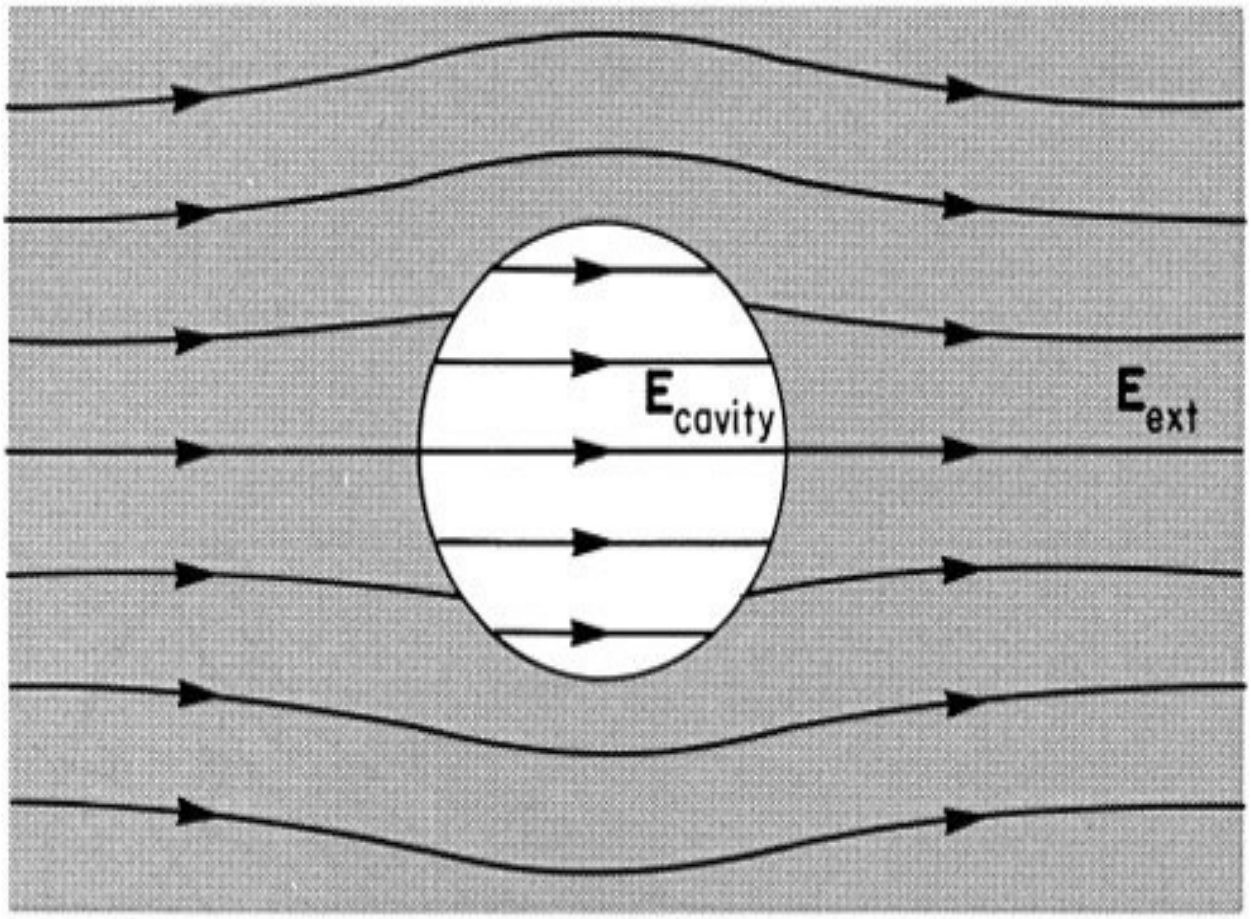
(3.8)

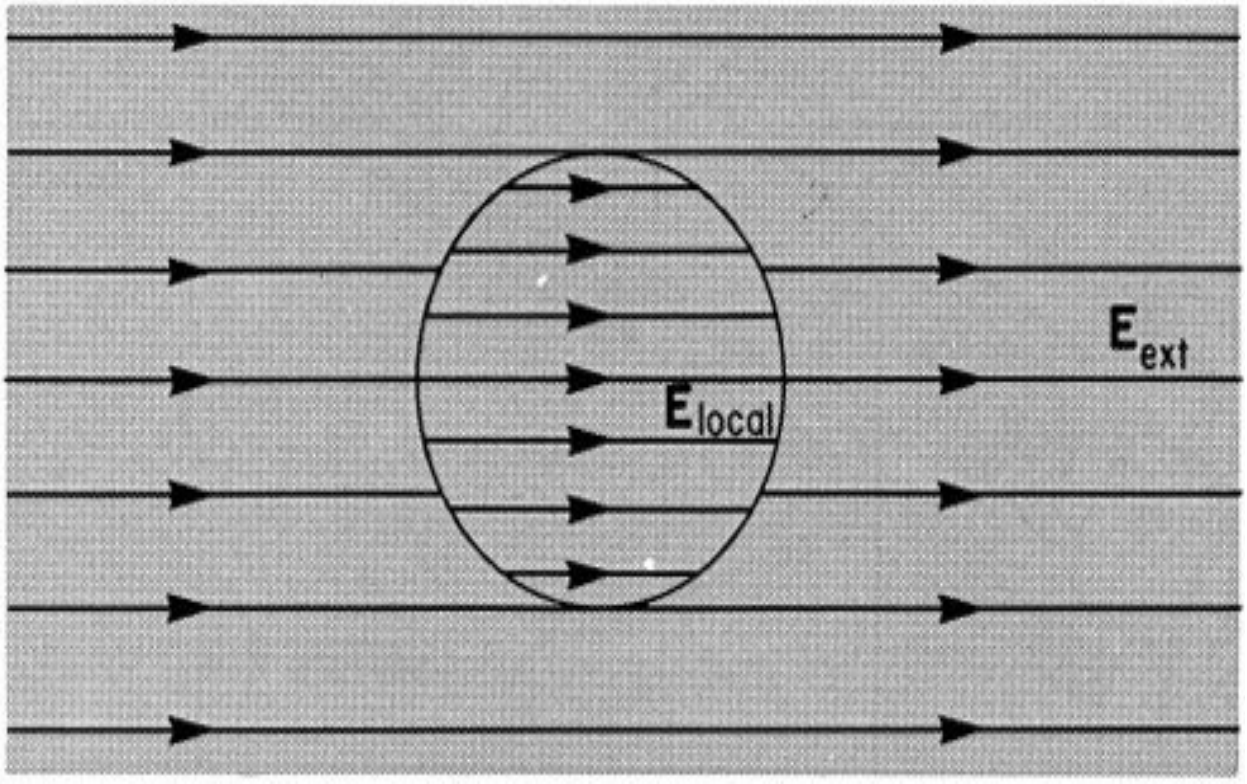
With this value of Eint, the potentials given by Equations (3.6) and (3.7) are continuous at the surface of the dielectric. It can easily be checked by differentiating these expressions that the potentials also satisfy the boundary conditions that $D_\perp$ and $E_\parallel$ are continuous across the surface. We have therefore found one possible form for the potential inside and outside the sphere, and the uniqueness theorem tells us that it is the only solution of Poisson's equation. (Again it must be emphasized that we have not systematically solved Poisson's equation in this example, but simply found the constants of integration after assuming the correct form of the solution.)

The relation between the internal field Eint, and the external field Eext is reminiscent of the Clausius–Mossotti formula for the local field in a liquid dielectric. We showed in section 2.2.3 that the local field is enhanced by a factor over the macroscopic field. In both cases the factor arises in the same way from the polarization charges of a uniformly-polarized material, induced over the surface of a sphere. But it must be pointed out that the local field calculated in the Clausius-Mossotti formula is not the same as the field in a spherical cavity. The local field is calculated from an atom placed in a uniformly polarized dielectric, and part of the field causing the polarization is due to the atom in question.

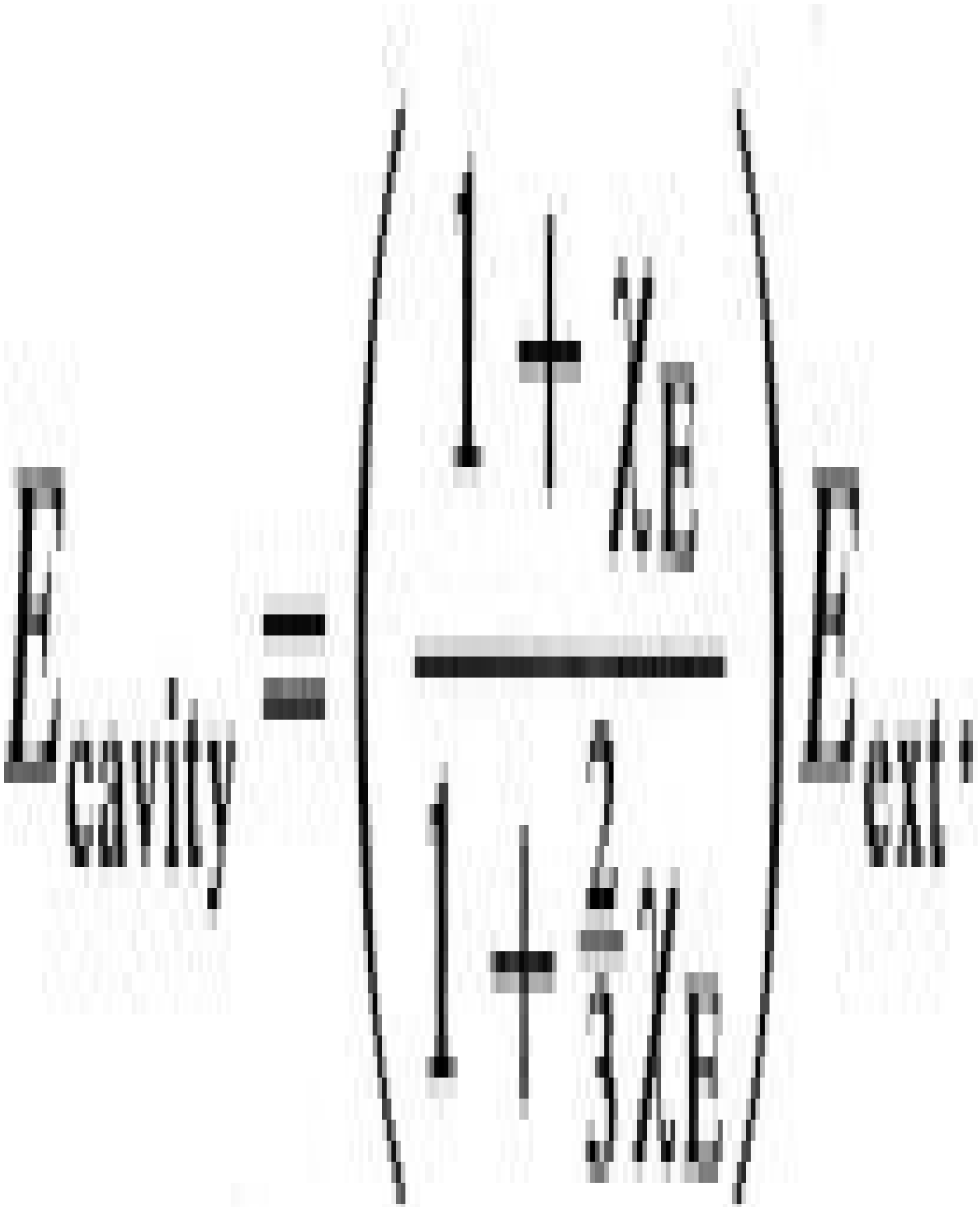
If the atom is not actually present in the dielectric, and there is a real cavity, the field and potential around the cavity are not uniform. The field in the cavity is less than the local field in a liquid which is uniformly polarized, as illustrated in Figure 3.7.

Figure 3.7. The distinction between the field in a cavity and the local field.

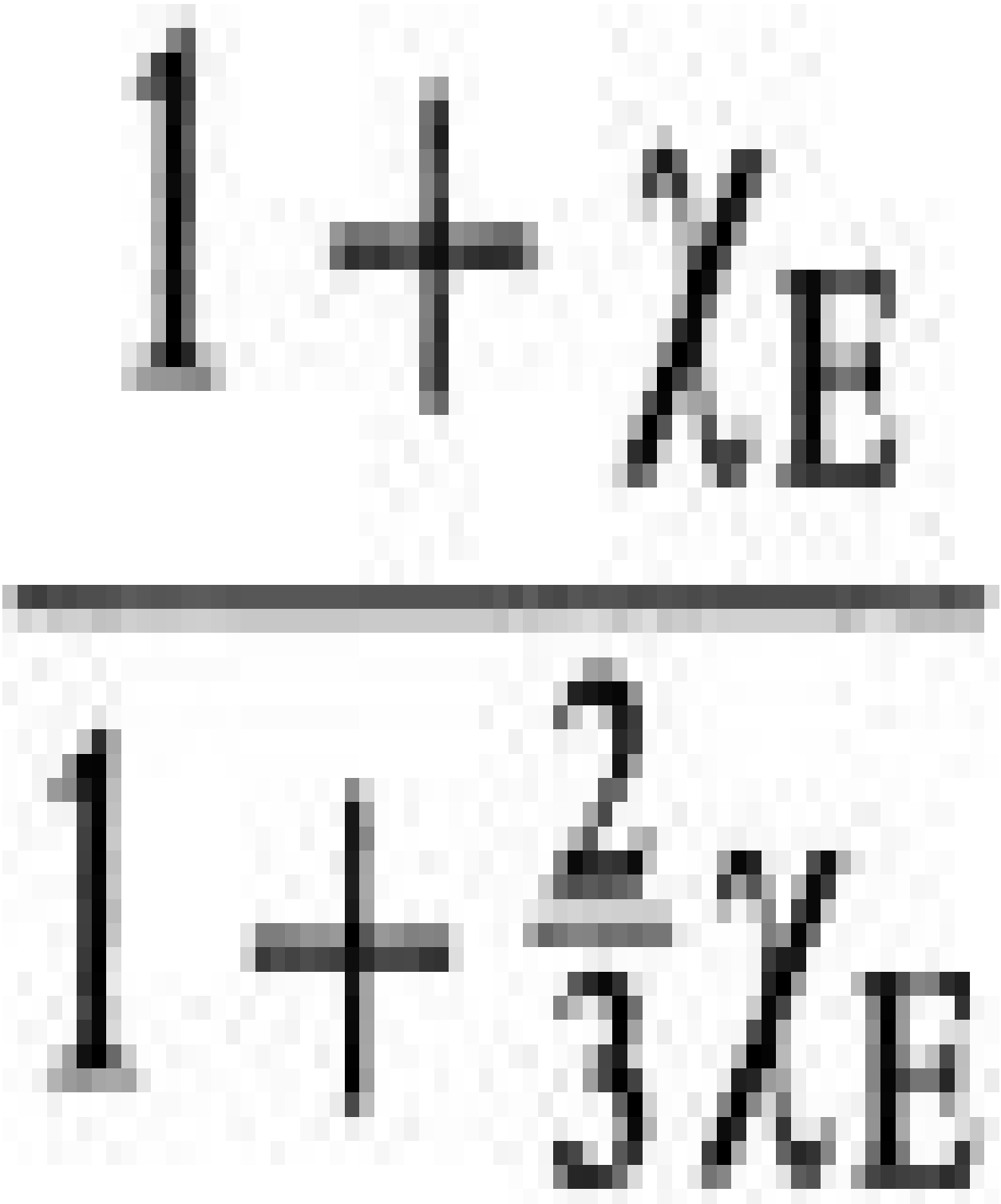

E_{cavity}
E_{ext}


E_{local}
E_{ext}

The field in a cavity is a little more awkward to work out than the field in a dielectric sphere, but again it is uniform and its value is



The enhancement factor

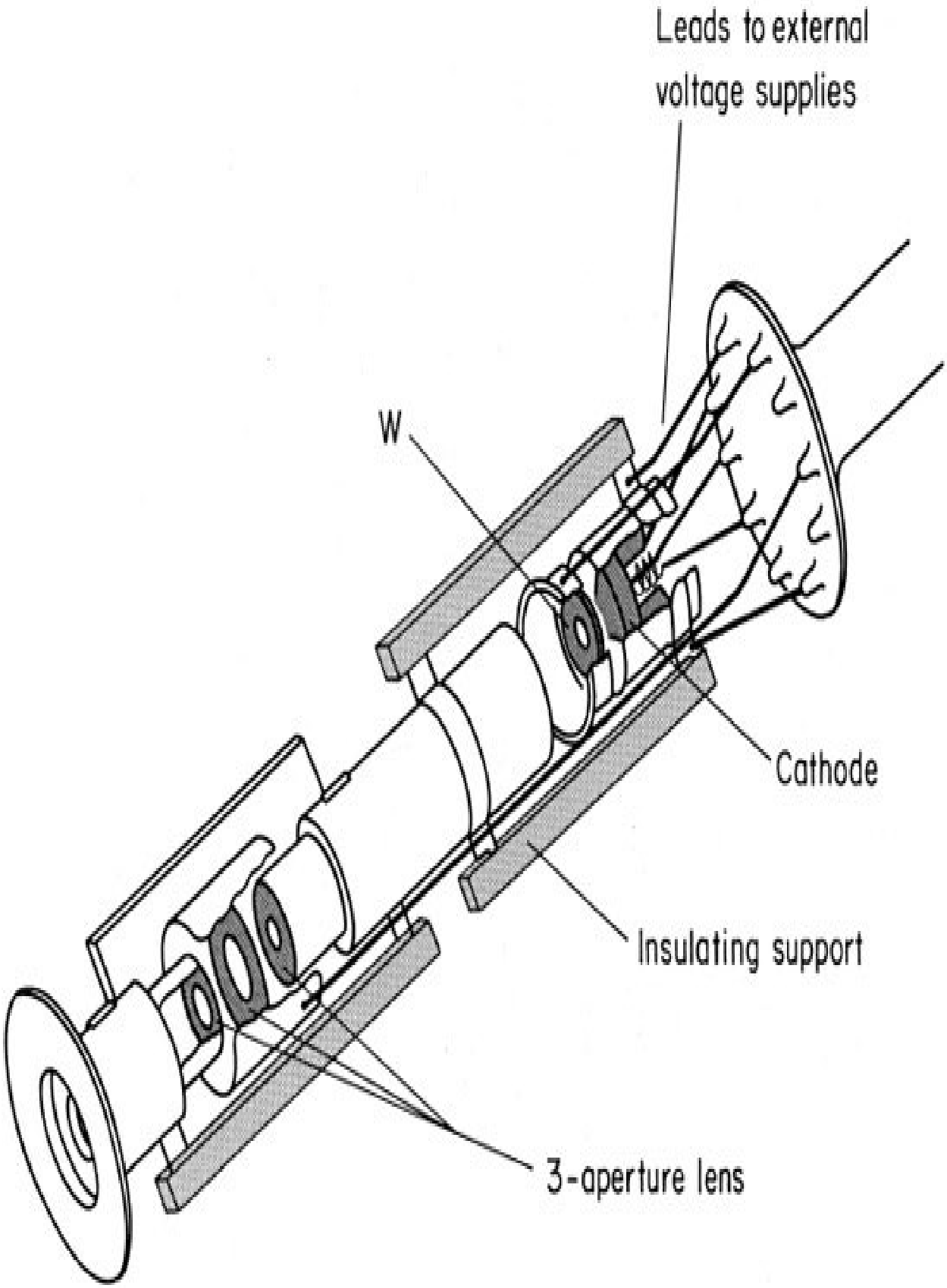


is not much different from when χ_E is small, because the presence of the cavity then causes only a small distortion of the neighbouring field.

3.4 ELECTROSTATIC LENSES

In most practical situations where one needs to know how the electric field varies with position, Laplace's equation has no exact solution in terms of simple functions. For example, the beam leaving the electron gun in a cathode ray tube passes a number of electrodes, as illustrated in Figure 3.8. The electrodes are made of conducting material, and each is maintained at a constant potential with respect to the cathode (the cathode is the heated electrode which is the source of the electrons). Around the electrodes there is a constant pattern of lines of force of the electric field. We need to know this pattern of lines of force before we can solve the equation of motion of an electron and find its path through the electron gun. In such a complicated geometry, there is no possibility of finding an exact solution to Laplace's equation, and we must be content with an approximation.

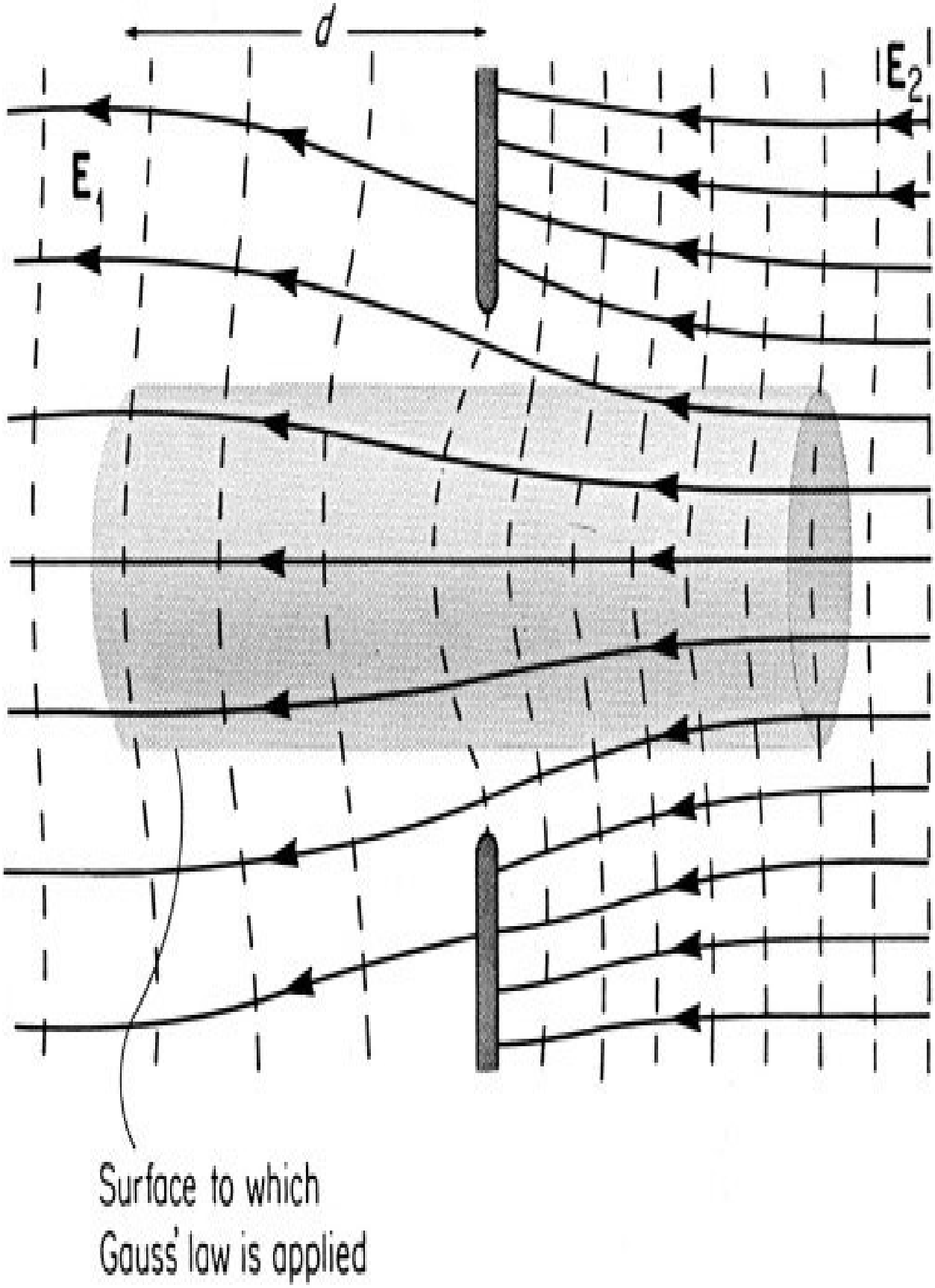
Figure 3.8. An electron gun.


Leads to external
voltage supplies
W
Cathode
Insulating support
3-aperture lens

Later on we shall briefly describe one of the methods of calculating approximate numerical values of the potential in complicated geometries. At present, however, we shall concentrate on the qualitative behaviour of the electron gun. We shall make some rather crude approximations, but as so often in physics, it is instructive to build up a rough picture of the way something works before worrying about the details in a special application.

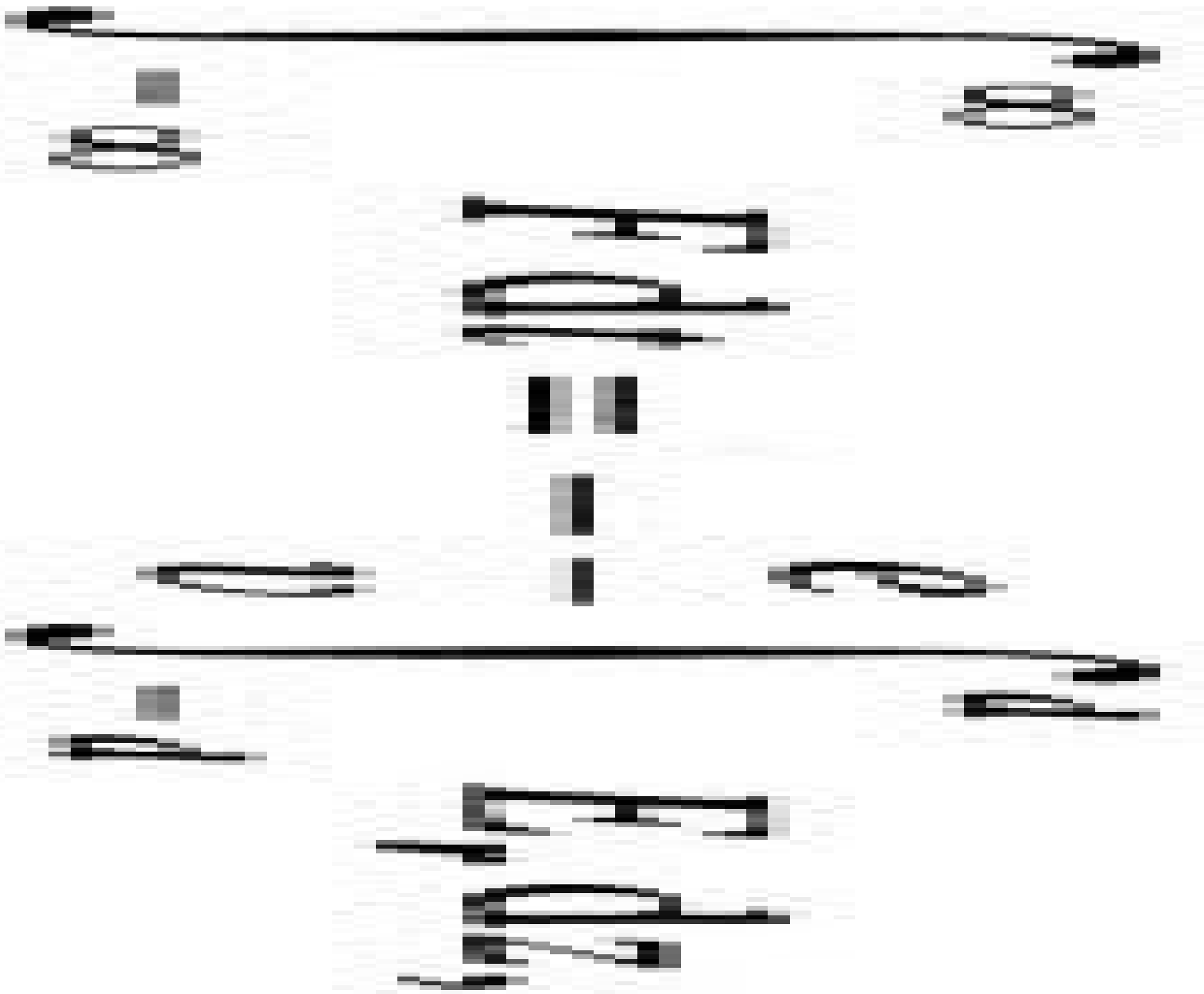
The drawing in Figure 3.8 was made from an actual electron gun used in an oscilloscope. The electrons leaving the cathode are already constricted into a narrow beam when they pass through the aperture in the electrode W, which is held at a negative potential with respect to the cathode. The remaining three electrodes are used to focus the beam onto a fine spot on the screen. Let us restrict attention to one of these focusing electrodes. What does the potential look like near its aperture? If we start by having an electrode with no aperture, with uniform but different fields on either side of it, then the equipotential surfaces are planes. Now punch a circular hole in the electrode. The electric field must change smoothly, and the field lines which are close together on the high field side of the aperture move apart on the low field side. The equipotential surfaces, always normal to the field lines, bulge through the aperture as shown in Figure 3.9.

Figure 3.9. The pattern of field lines and equipotentials near an aperture.


d
E1
E2
Surface to which
Gauss' law is applied

Near the aperture, the electric field has a radial component directed away from the axis. An electron, arriving at the aperture on a path parallel to the axis, experiences a force deflecting it towards the axis. There is also an electric field along the axis, but in practical electron guns the electron already has an energy of about one keV when it reaches the focusing electrodes, and it is not a bad approximation to assume that the axial velocity of the electron is constant. We only need to consider the radial field when estimating the deflection of the electron.

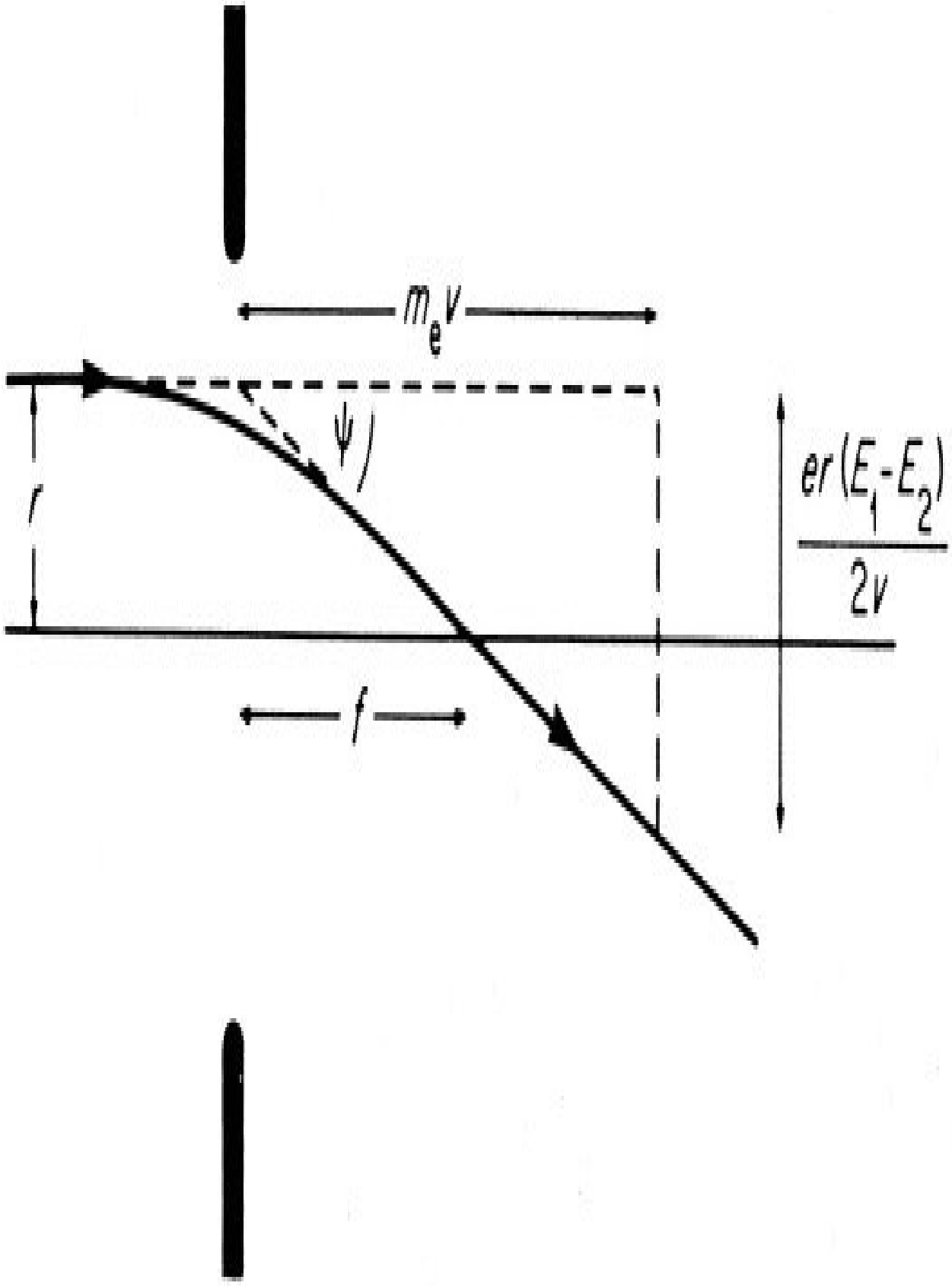
Let us follow the trajectory of an electron which approaches parallel to the axis, and is then deflected towards the axis as it passes through the aperture. Such a trajectory is illustrated by the thick line in Figure 3.10, for an electron initially at a distance r from the axis. We shall assume that the radial field is negligible at points further away than a distance d from the aperture. Taking the origin of cylindrical coordinates at the centre of the aperture, the electron is in a radial field Er between the planes $z = \pm d$, and it then experiences a radial force $F = -eE_r$; the minus sign indicates that the force is directed towards the axis. An electron with axial velocity v moves a distance dz in a time $dt = dz/v$, and acquires a radial momentum $Fdt = -eE_r(dz/v)$. The total radial momentum acquired by the electron in passing through the aperture is



neglecting the change of the electron's axial velocity as it moves through the aperture. Provided that the deflection is small, we can estimate the integral on the right-hand side of this equation, without knowing the detailed form of Er, by applying Gauss' theorem to the cylindrical surface of radius r and length 2d shown shaded in Figure 3.9.

At the end of the cylinder, the field is uniform and pointing along the axis, with magnitudes E1 and E2 on either side of the aperture. The outward flux at the two flat surfaces is therefore $\pi r^2(E1 – E2)$. Only the radial component Er contributes to the flux out of the curved surface: the area of a circular strip of length dz on the surface is $2\pi r\, dz$, and the total flux out of the curved total flux out of the curved surface is . The whole cylindrical surface encloses no charge, and according to Gauss' law the total outward flux is zero;

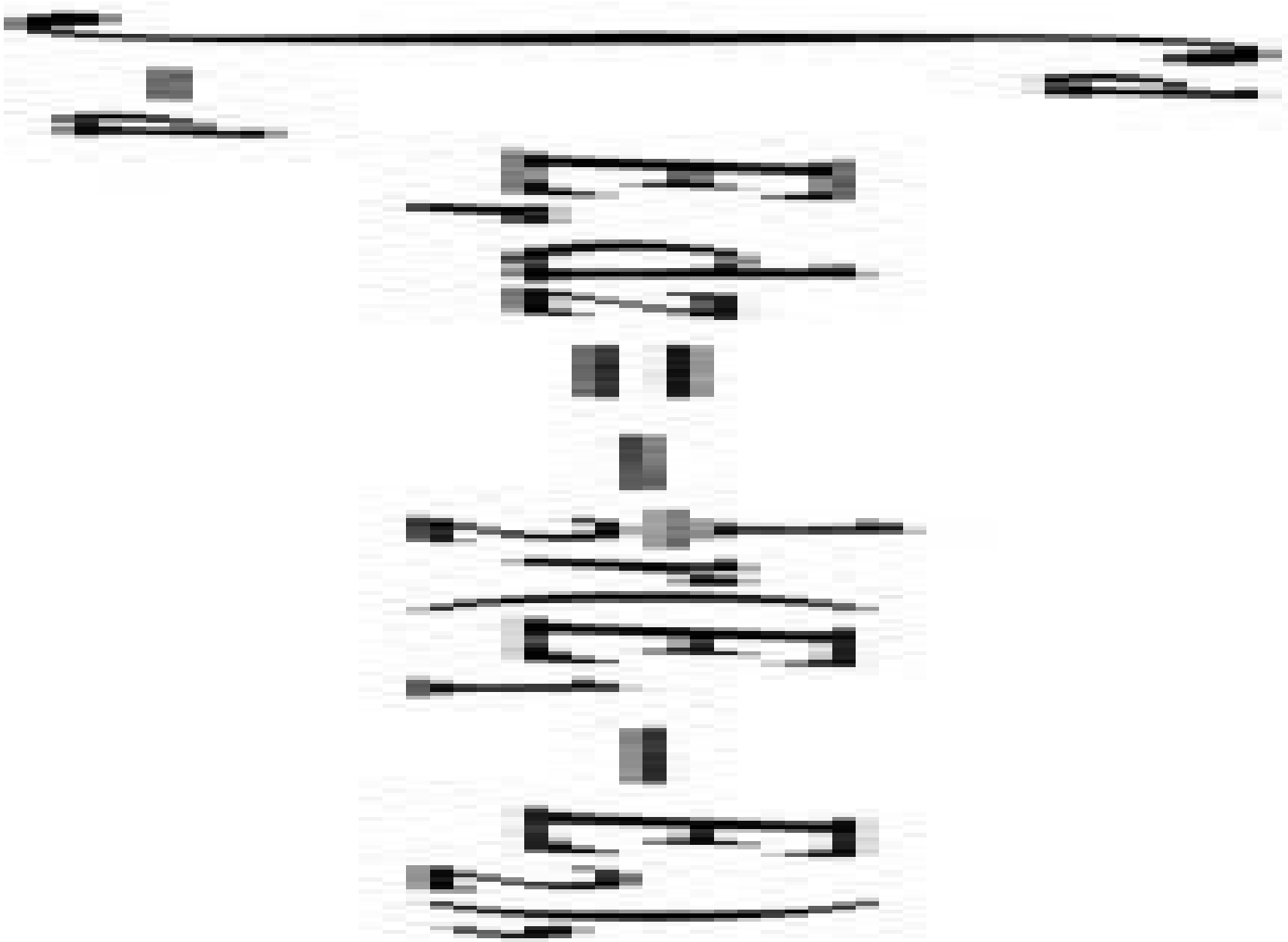
Figure 3.10. An electron trajectory through the aperture.


$m_e v$
r
ψ
$\frac{er(E_1-E_2)}{2v}$
f

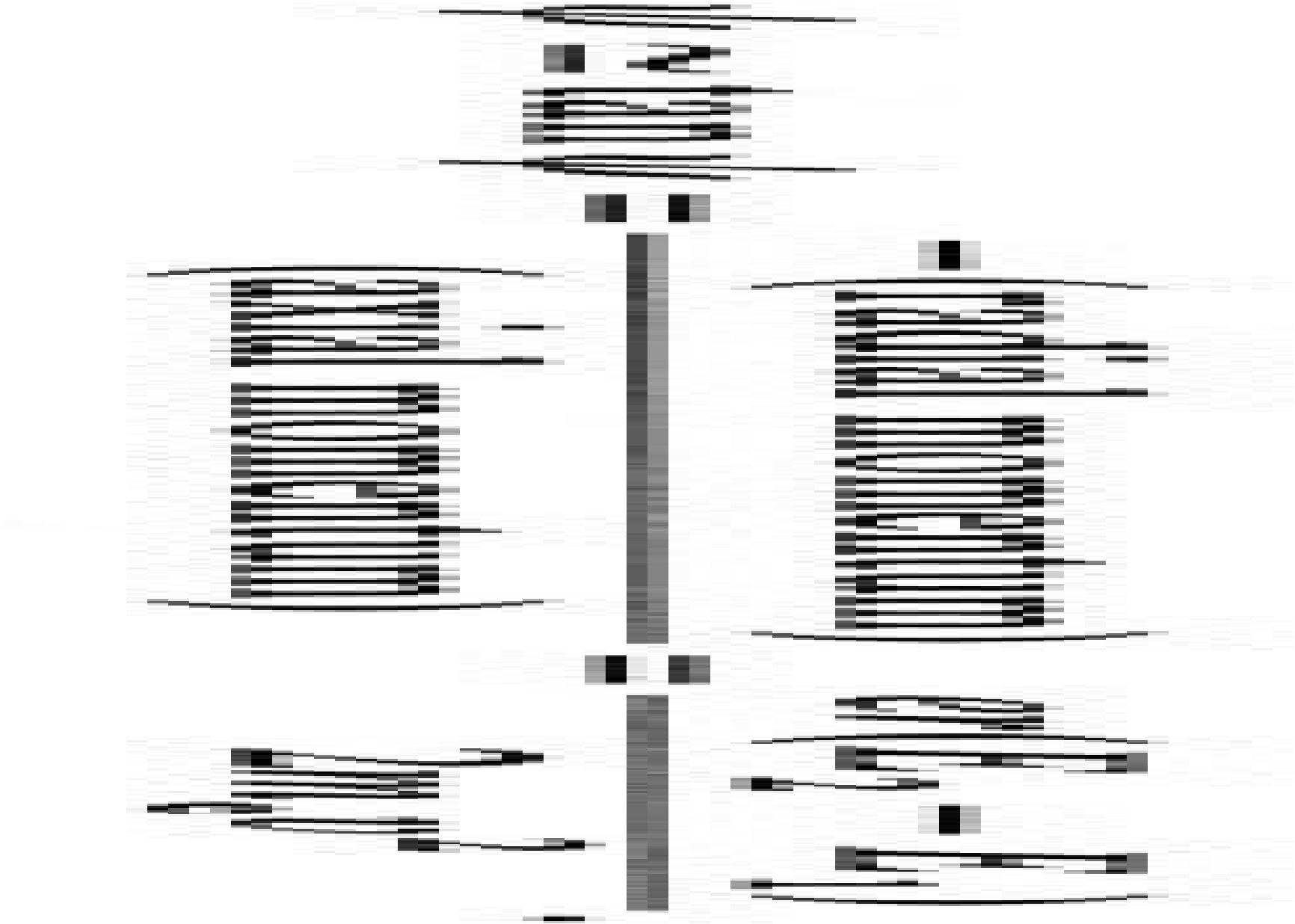
i.e.

$$\pi r^2(E_1 - E_2) + 2\pi r \int_{-d}^{d} E_r \, dz = 0.$$

Hence

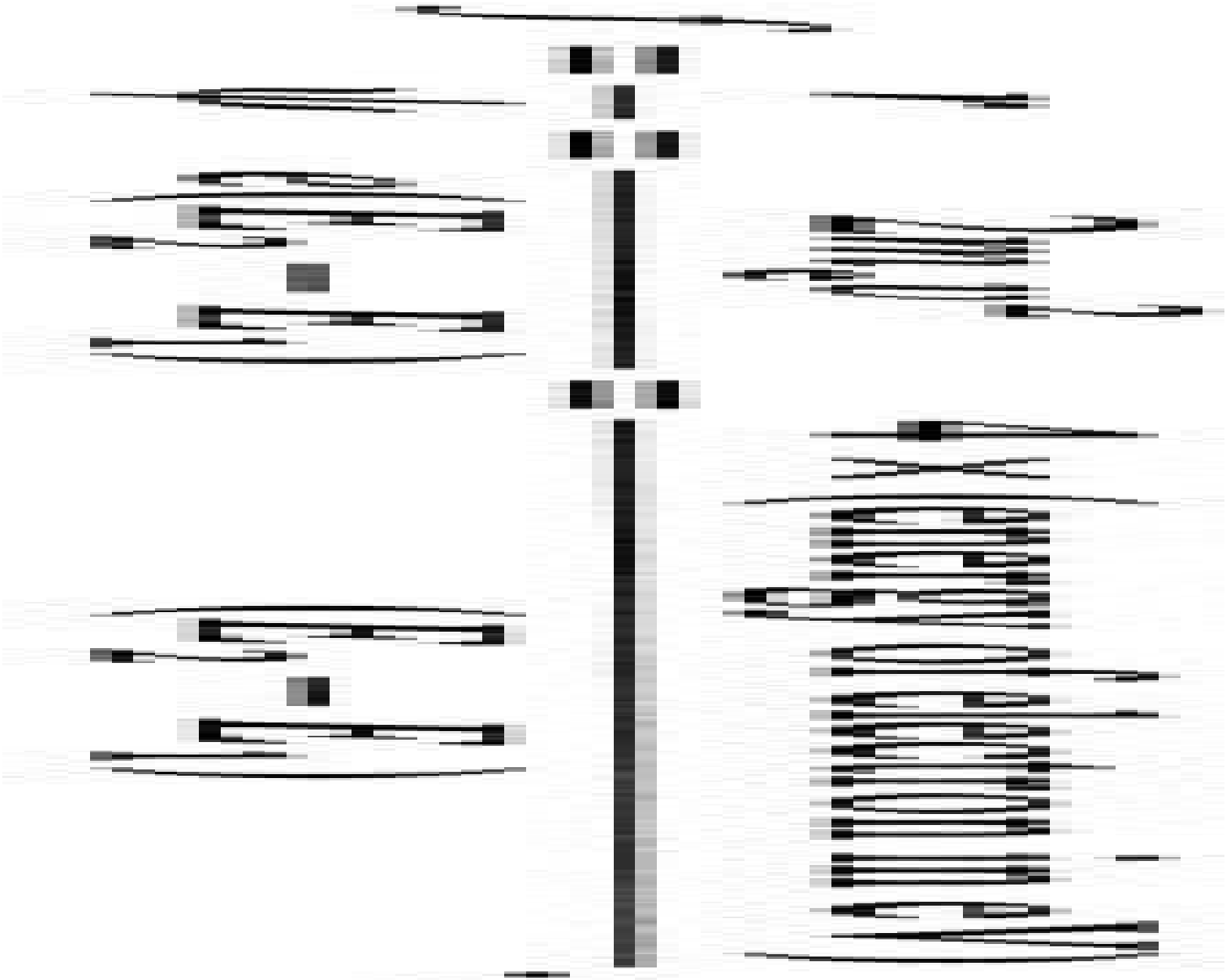


and the radial momentum acquired by the electron in passing through the aperture is + er(E1 − E2)/2v. The angle of deflection ψ of the electron towards



In optics, when a ray passes through a thin lens of local length f, at a distance r from the axis, the ray is deflected through an angle $\psi = r/f$. Because the electron deflection in passing through the aperture is also proportional to r, the aperture is acting like a lens of focal length

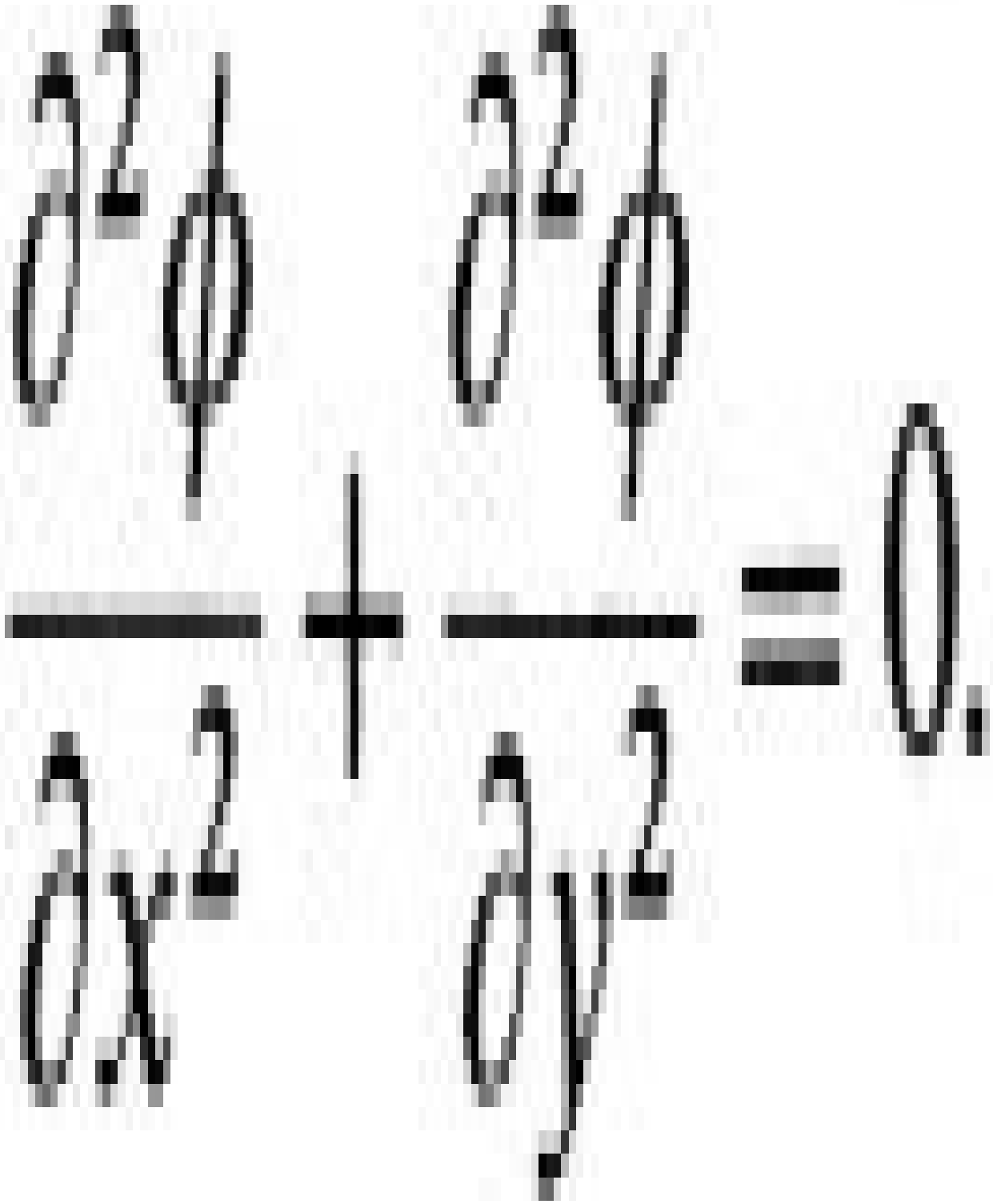
(3.9)



A single aperture therefore acts as an electron lens. However, just as cameras are made with compound optical lenses, so an electron gun usually has a number of focusing apertures. A set of three apertures as in Figure 3.8 is commonly used: one advantage of this arrangement is that there can be field-free regions on either side of the three-aperture-lens, with the outer electrodes kept at a fixed potential. The focal length of the lens can be varied simply by changing the potential of the inner electrode. We can make a rough estimate of the performance of the lens in Figure 3.8 by assuming it to be made up of three thin lenses with focal lengths given by Equation (3.9). Suppose that the separation of each pair of electrodes is 1 cm, that the outer electrodes are both at a potential of 1 kV and that the inner electrode is held at 250 V. Application of the thin lens formula to each aperture in turn shows that IkeV electrons travelling parallel to the axis are brought to a focus 12 cm from the central electrode, i.e. the three-aperture system acts as a lens with focal length 12 cm.

3.5 NUMERICAL SOLUTIONS OF POISSON'S EQUATION

For many practical problems in which the electrostatic field needs to be known, there is no analytical solution to Poisson's equation, and all that we can do is to generate numerical solutions. The equation is a second order differential equation in one, two or three dimensions. It must be solved subject to boundary conditions which define the geometry of the problem and specify fields or potentials at the boundaries, with perhaps some additional conditions relating to charges on surfaces or inside the region. There are many different numerical techniques for solving differential equations, most developed for digital computers. A method which is of very general applicability is the finite difference method, in which the value of the potential is calculated at a large number of points on a mesh spanning the region of interest. Approximate values of the derivatives can be found from the differences of potential at neighbouring points, and the differential equation converted to a relation approximately satisfied by the values of the potential at these points. We shall illustrate the systematic procedure leading to an approximate numerical solution by considering a problem in which there is no free charge present, and in which there is no variation in the potential in one direction, which we shall take to be the z-direction. The equation to be solved is then Laplace's equation, and in Cartesian coordinates it reduces to



We shall consider a square region in the x–y plane, bounded on two sides by equipotentials at 0 and 10 volts, and on the other two sides by boundaries along which the potential changes linearly. (The potential obviously varies linearly everywhere between the equipotentials, but this trivial problem serves to illustrate the numerical techniques.) Now mark out a grid of points in the region where we want to know the potential, as in Figure 3.11. Each point is a distance h from its four nearest neighbours. Concentrate attention for the moment on the point labelled G. If the grid is fine enough, it will be a good approximation to express the potential at the neighbouring points C, F, H and K as a Taylor series about the potential ϕG at G, retaining only the second-order terms. Thus

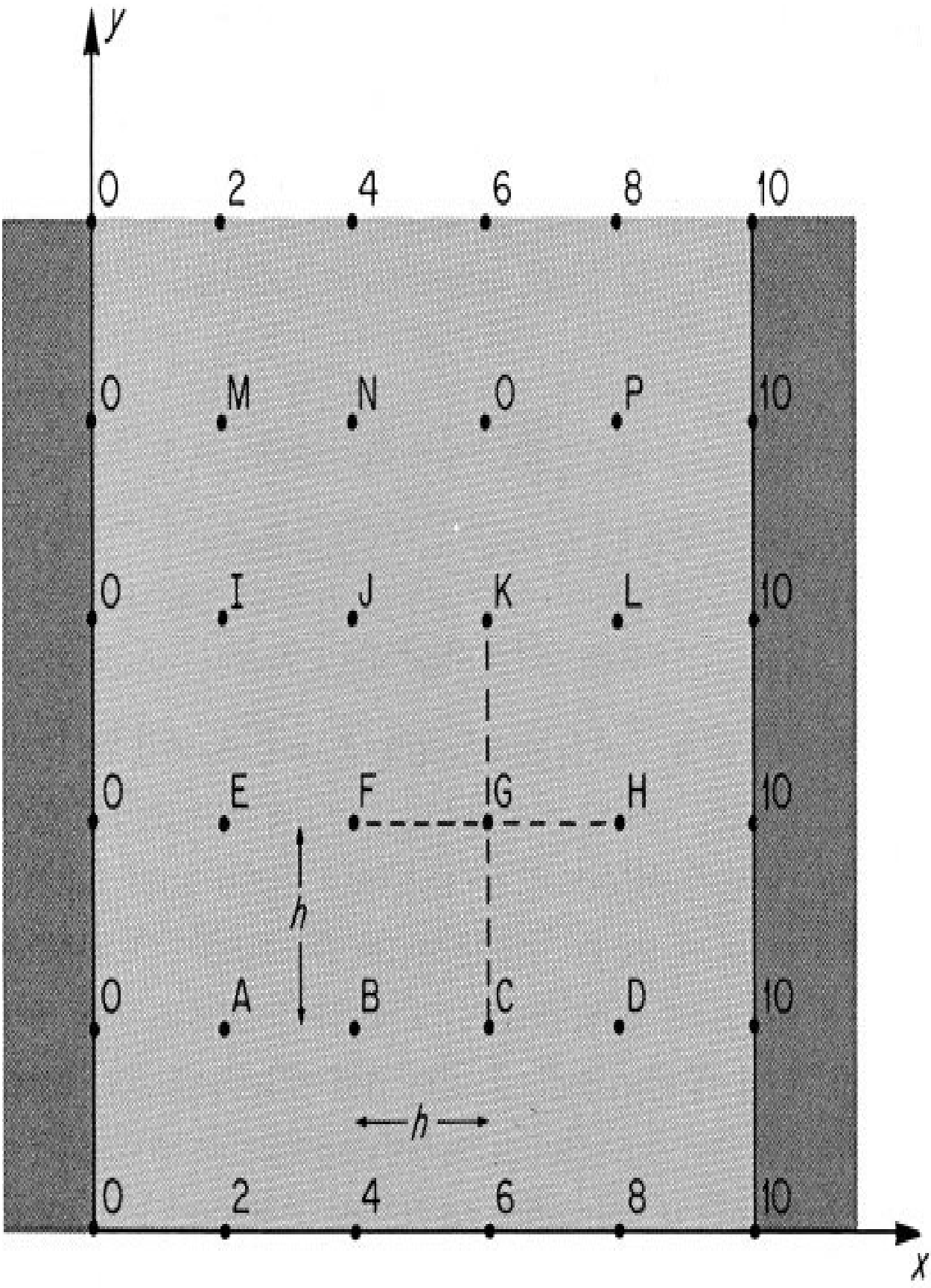
$$\phi_F = \phi_G - h\left.\frac{\partial \phi}{\partial x}\right|_G + \tfrac{1}{2}h^2\left.\frac{\partial^2 \phi}{\partial x^2}\right|_G + \cdots$$

$$\phi_H = \phi_G + h\left.\frac{\partial \phi}{\partial x}\right|_G + \tfrac{1}{2}h^2\left.\frac{\partial^2 \phi}{\partial x^2}\right|_G + \cdots$$

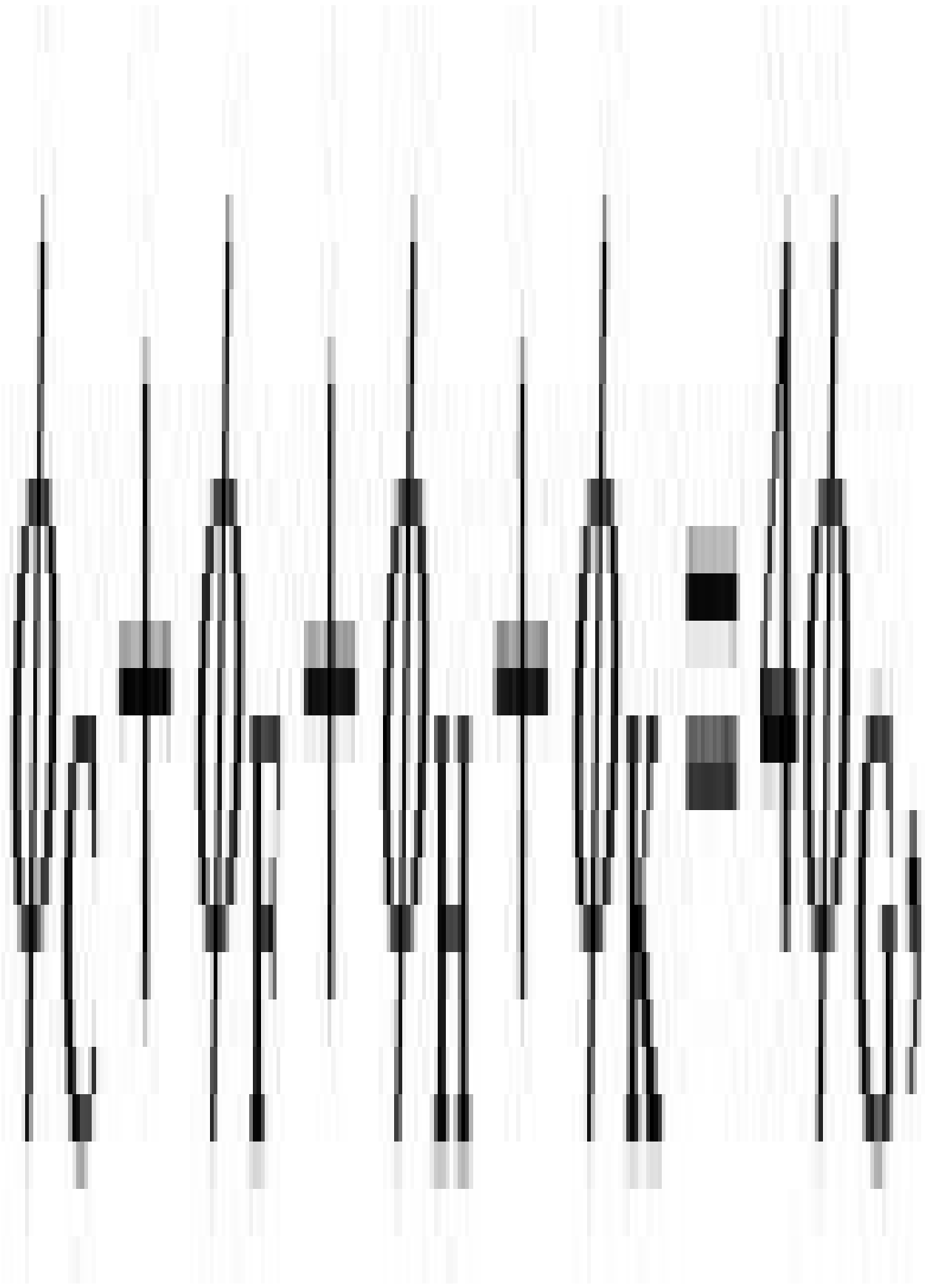
$$\phi_C = \phi_G - h\left.\frac{\partial \phi}{\partial y}\right|_G + \tfrac{1}{2}h^2\left.\frac{\partial^2 \phi}{\partial y^2}\right|_G + \cdots$$

$$\phi_K = \phi_G + h\left.\frac{\partial \phi}{\partial y}\right|_G + \tfrac{1}{2}h^2\left.\frac{\partial^2 \phi}{\partial y^2}\right|_G + \cdots.$$

Figure 3.11. An equally spaced mesh of points used in finding an approximate solution to Laplace's equation.


y
0
2
4
6
8
10
M
N
O
P
I
J
K
L
E
F
G
H
A
B
C
D
h
h
x

Adding the four equations, and using Laplace's equation to eliminate the second order terms, we find

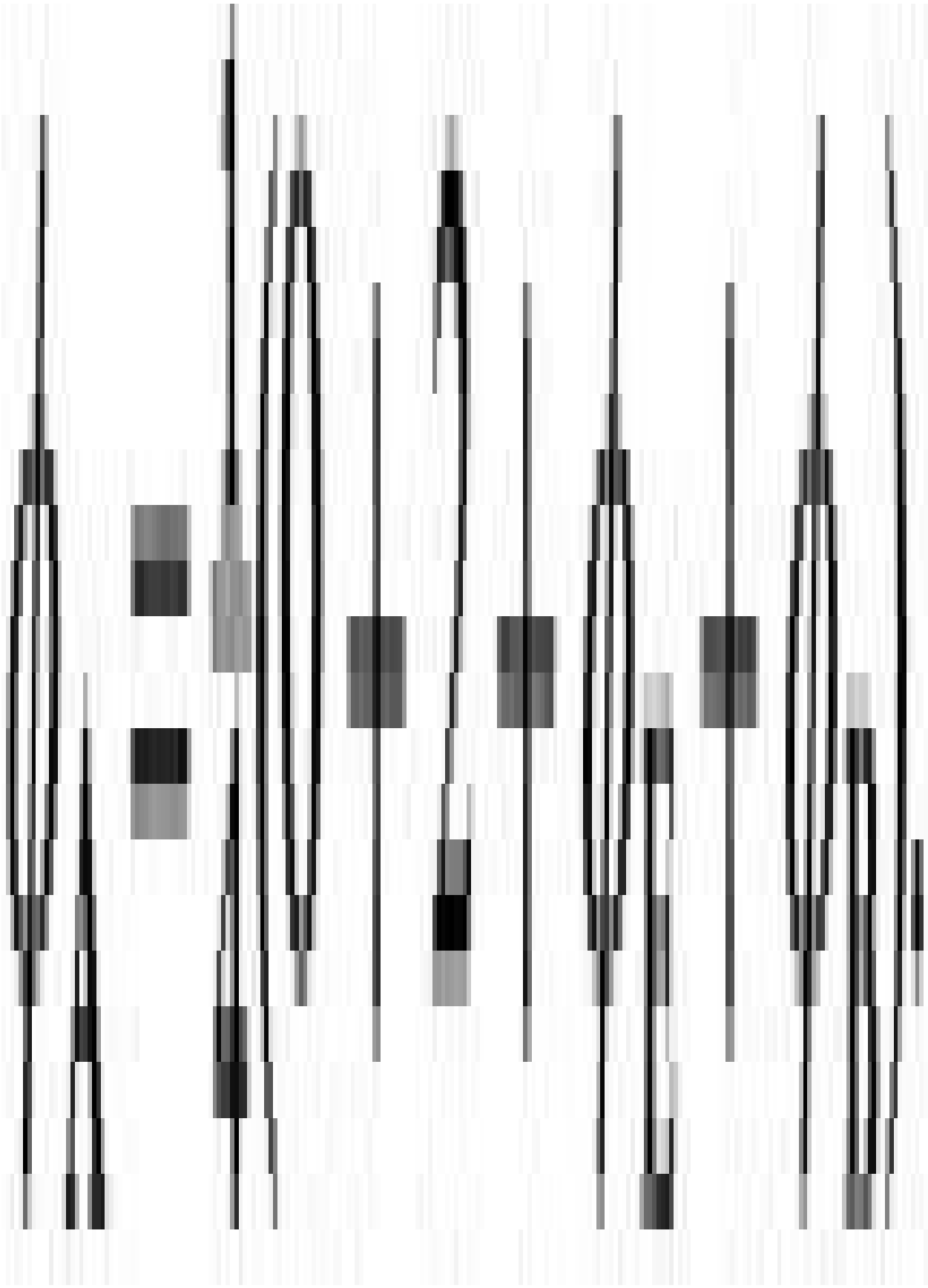


or

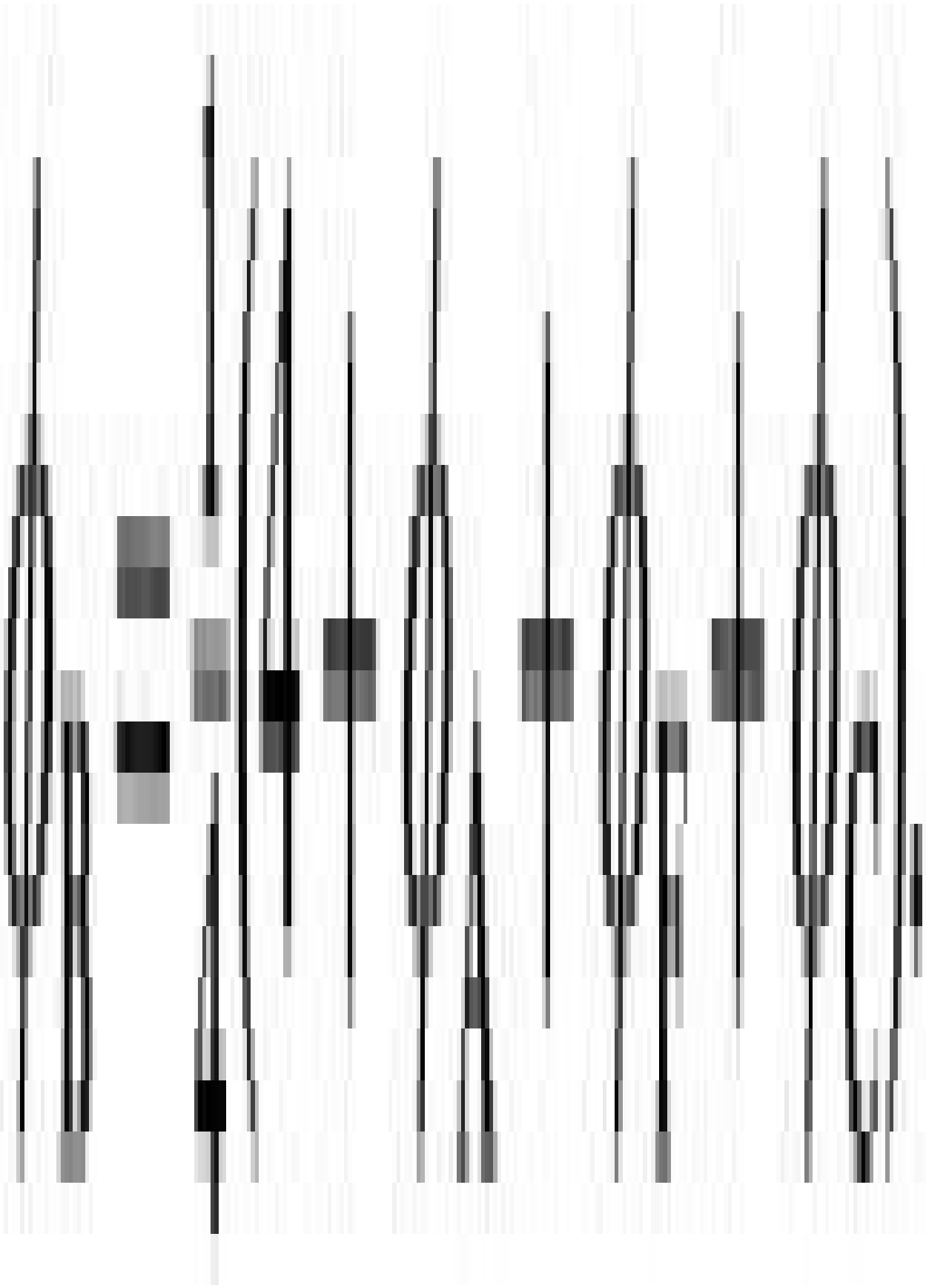
(3.10)

This equation is not exact, because we have ignored the third and higher order terms in the Taylor expansion. However, these terms can be made as small as we please by choosing the grid sufficiently fine.

Now suppose that we can find a set of values for the potential at the grid points which satisfy the boundary conditions, and for which each value is the mean of the four nearest neighbour values, as required by Equation (3.10). Such a set of values is a good approximation to the exact potential at the points of the grid. To find this set of values, we must solve a large number of simultaneous equations —in fact there is one equation of the form 3.10 for each point of the grid. These equations can be solved by an iterative procedure. One starts with a guess at the potential distribution. For example, if in the grid in Figure 3.11 the values of the potential on the boundary are as labelled, the exact solution is obviously a potential varying linearly from 0 to 10 volts from left to right. Let us make a rather poor initial guess at this potential, assuming it to be zero everywhere except on the boundary. A lot of the points on the grid do not now satisfy Equation (3.10). However, we can systematically adjust them, starting for example at point A. The first guess for the potential at A is $\phi A = 0$, but a better guess is



Putting $\phi E = 0$ and $\phi B = 0$ leads to $\phi A = 0.5$. Similarly



Using the new guess of $\phi A = 0.5$, and putting $\phi F = 0$, $\phi C = 0$, we find $\phi B = 1.125$.

After evaluating the new approximation to the potential at each point, the whole operation is carried out over and over again until Equation (3.10) is satisfied at every point on the grid. In the present example, the method converges to within 1% of the exact solution only after 18 iterations. The method is thus very laborious, and it also suffers from the disadvantage, common to all numerical methods, that the solution is only applicable to one set of boundary conditions. If some of the boundary potentials in Figure 3.11 are changed, the whole iterative procedure has to be carried out again. Nevertheless finite difference methods are useful because of the following features:

(i) the potential is evaluated in a systematic manner for an arbitrary set of boundary conditions.

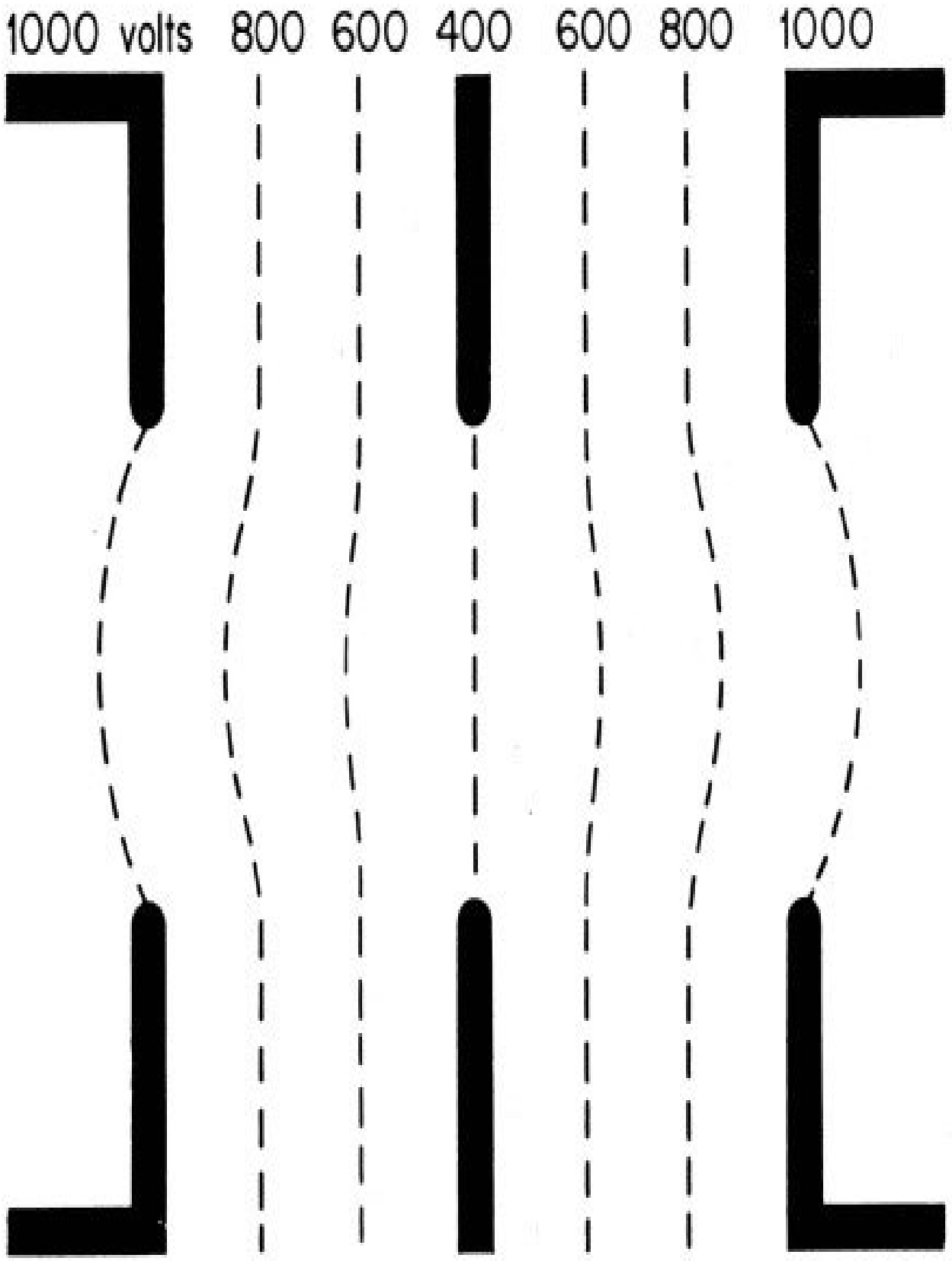
(ii) each step in the calculation is straightforward, and although the iteration may be lengthy, it is ideally suited to calculation on a digital computer.

Finite difference methods can be applied to problems in three dimensions, or to problems in which Cartesian coordinates are not appropriate. In all cases, a Taylor expansion leads to approximate relations between the potentials at neighbouring points (though the relations are not always as simple as Equation (3.10). This technique of dividing up a region into a finite grid is nowadays the most important method of finding approximate solutions to Laplace's equation, since digital computers are fast enough, and have enough capacity, to cope with the large number of variables which are needed in two or three dimensions.

A finite difference method has been used to compute the potential distribution around the three aperture lens of an electron gun, in which the separation of the

electrodes is equal to the diameter of the apertures*. When the outer electrodes are at l kV, this lens system has a focal length of 12 cm when the inner electrode is at 400 volts. The value of 250 volts, derived earlier using the thin lens formulae, is a poor approximation for the practical lens with wide apertures. The potential distribution is drawn in Figure 3.12.

Figure 3.12. The equipotentials in a three-aperture lens.


1000 volts
800
600
400
600
800
1000

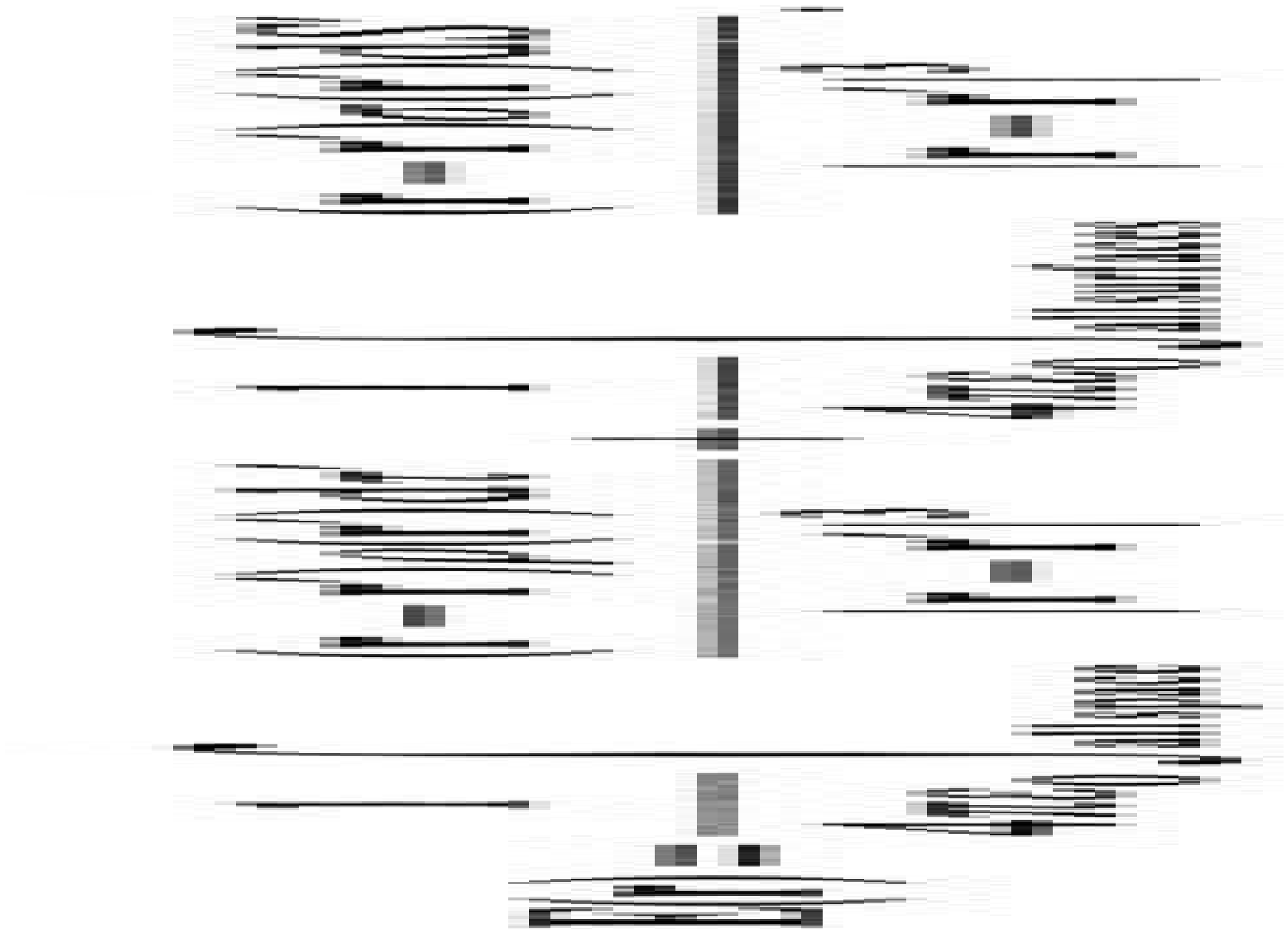
3.6 SUMMARY OF ELECTROSTATICS

The key definitions and relations of electrostatics are set out here, numbered as in their first appearance in the text.

Coulomb's law states that the force between two charges q1 and q2 is inversely proportional to the square of the distance between them

(1.2)

The electric field at a point with position vector r is the force per unit charge on an imaginary (and vanishingly small) test charge placed at r.



(1.10)

Gauss' law states that the flux of E out of a closed surface S enclosing a volume V is (l/ε0) times the total charge within V

(1.14b)

Gauss' law can be expressed equivalently in differential form by the equation

(1.20)

The potential difference between two points A and B is defined as the work done per unit charge in moving a small test charge from A to B:

(1.24)

The differential form of this equation is

(1.26)

The electric field is conservative, which means that no work is done in taking a test charge around any closed loop, i.e.

(1.23)

The potential energy of an assembly of point charges qi is the work done in bringing them up one by one from infinity to their final positions. Choosing the potential to be zero at infinity, and calling ϕi the potential at qi due to all the other charges, the potential energy is

(1.36)

Potential energy can also be expressed in terms of the field by associating an energy density $\varepsilon 0E^2$ with the field. The total energy of a system of charges generating a field E is

(1.42)

The capacitance of a pair of conductors is the ratio of the charge stored on them to the voltage difference between them:

(1.39)

The capacitance of a pair of conductors is free space is determined only by their geometry. When the space between the conductors is completely filled with a homogeneous dielectric material, the capacitance increases by a factor ε, the relative permittivity of the material. The increase in capacitance arises because the dielectric material becomes polarized and in a field E it acquires a dipole

moment per unit volume, or polarization

(2.5)

The factor $(\varepsilon - 1)$ has a special name. It is called the electric susceptibility:

(2.7)

When discussing dielectric materials, it is useful to introduce a new vector field, the electric displacement D defined by

(2.20)

The electric displacement satisfies a relation rather like Gauss' theorem for the field E, but referring only to the free charge density ρ_f:

(2.19)

At boundaries between regions containing materials of different relative permittivity the. boundary conditions are

(2.23)

(2.24)

Within a region containing uniform dielectric with relative permittivity ε, the potential satisfies Poisson's equation:

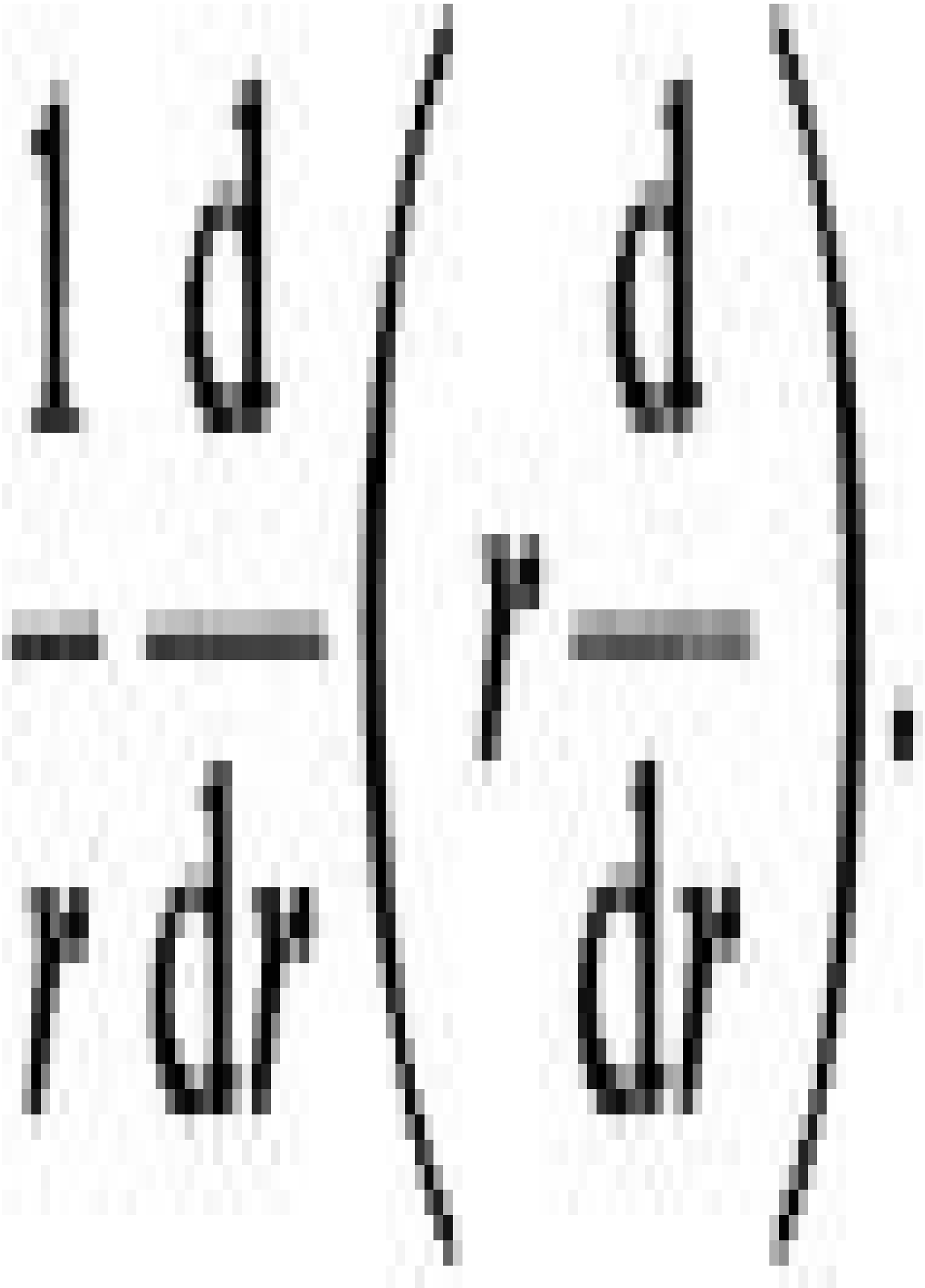
(3.3)

Where no free charges are present, Poisson's equation reduces to Laplace's equation:

(3.4)

PROBLEMS 3

3.1 An electron beam has a circular cross-section of radius R, and the charge density within the beam has the constant value ($-\rho_0$). Solve Poisson's equation to find the electric field. In cylindrical coordinates the radial part of the Laplacian operator is



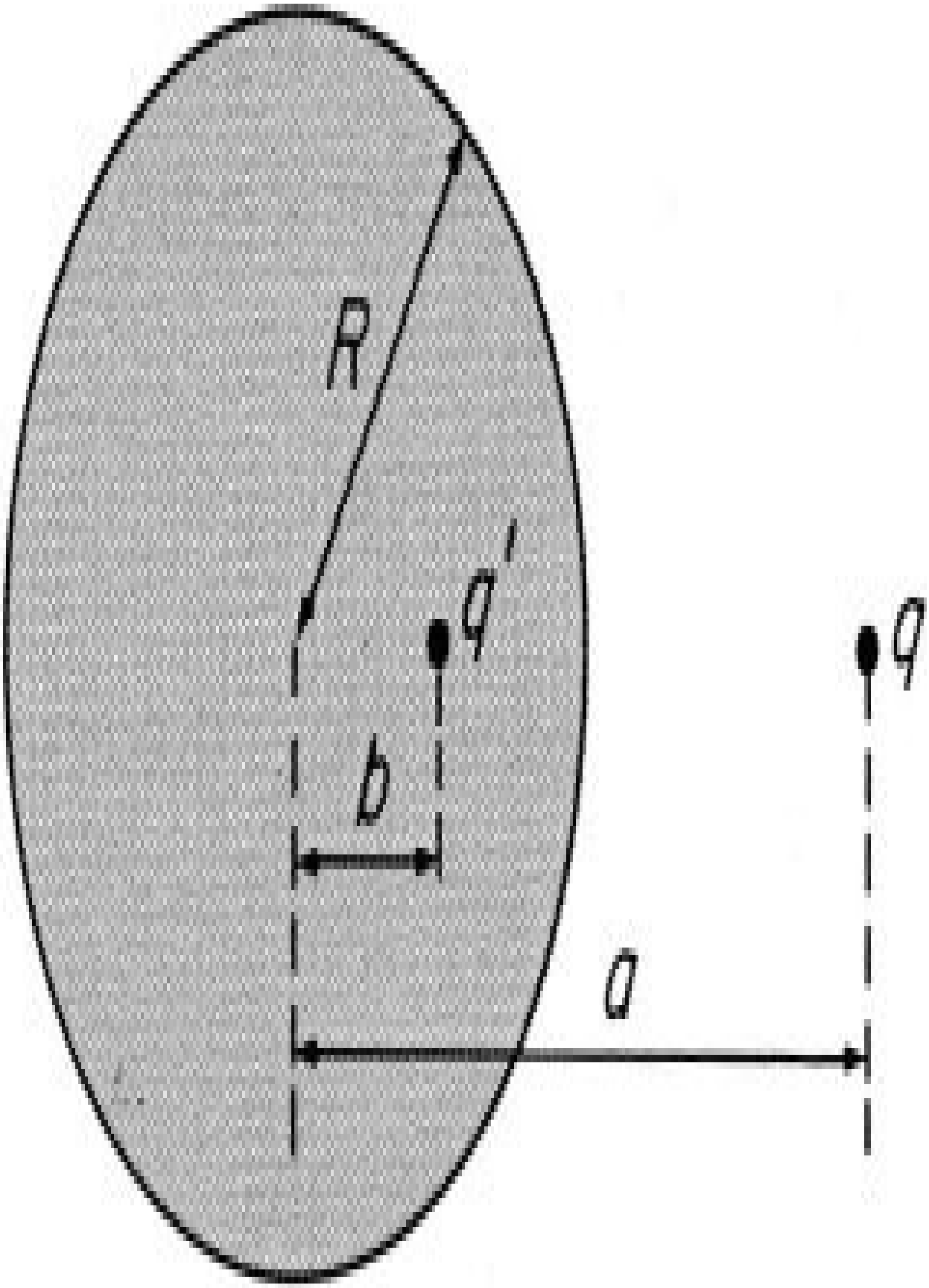
3.2 The potential distribution in a thermionic diode is quite different from that in a silicon diode. The space between the electrodes is evacuated, and if the electrodes are parallel plates, the field between them is nearly uniform when the electron density in the diode is small, as occurs when the anode is at a high potential with respect to the cathode. But at low voltages, electrons move slowly near the cathode, and their charge density becomes high: the diode becomes space-charge limited when the electric field at the cathode is zero. Find the current density in a space-charge limited parallel plate diode with plate separation d when there is a voltage V across the plates. (Tackle the problem in the following way: (i) express the electron velocity in terms of the potential between the plates; (ii) this gives the charge density in terms of the current density; (iii) solve Poisson's equation with the conditions $\phi = 0$ and V at the electrodes, $\partial\phi/\partial x = 0$ at the cathode.)

3.3 A proton is at rest 1 cm below a conducting plate. Does the electrostatic attraction towards the plate overcome the gravitational force acting on the proton? (Hint: what image charge can reproduce the surface of the plate as an equipotential?)

3.4 A telephone wire of diameter 1 mm is suspended 10 m above the ground. Estimate its capacitance to earth per unit length. (Images again.)

3.5 A point charge q is situated at a distance a from the centre of a conducting sphere of radius R as shown in Figure 3.13. Show that the image charge q' satisfies the condition that the surface of the sphere be an equipotential if $q' = -qR/a$ and $b = R^2/a$. What is the force attracting q towards the sphere if the sphere carries no net charge?

Figure 3.13.

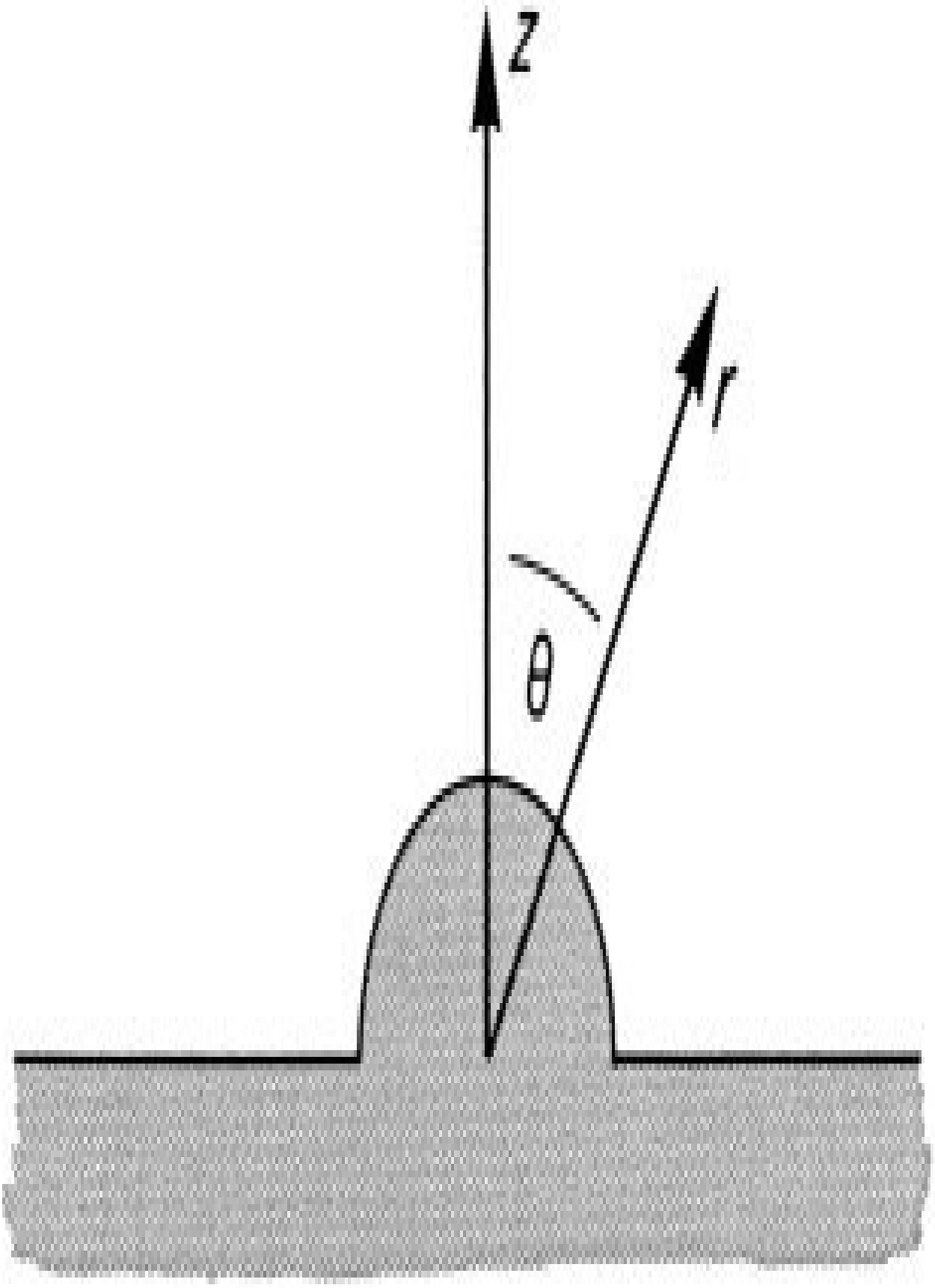

R
q'
q
b
a

3.6 Two conducting spheres, each of radius 2 cm have their centres 10 cm apart. If the breakdown field in air is 30kV/cm, estimate the breakdown voltage across the sphere gap. (Sphere gaps of this kind are often used as an overvoltage protection in high voltage equipment.)

3.7 A sheet of charge of density σ/unit area lies midway between two earthed conducting plates. What are the surface charge densities induced on the plates? (You cannot use images for this one, because an infinite number of sheets of image charge are required. Use Gauss' law.)

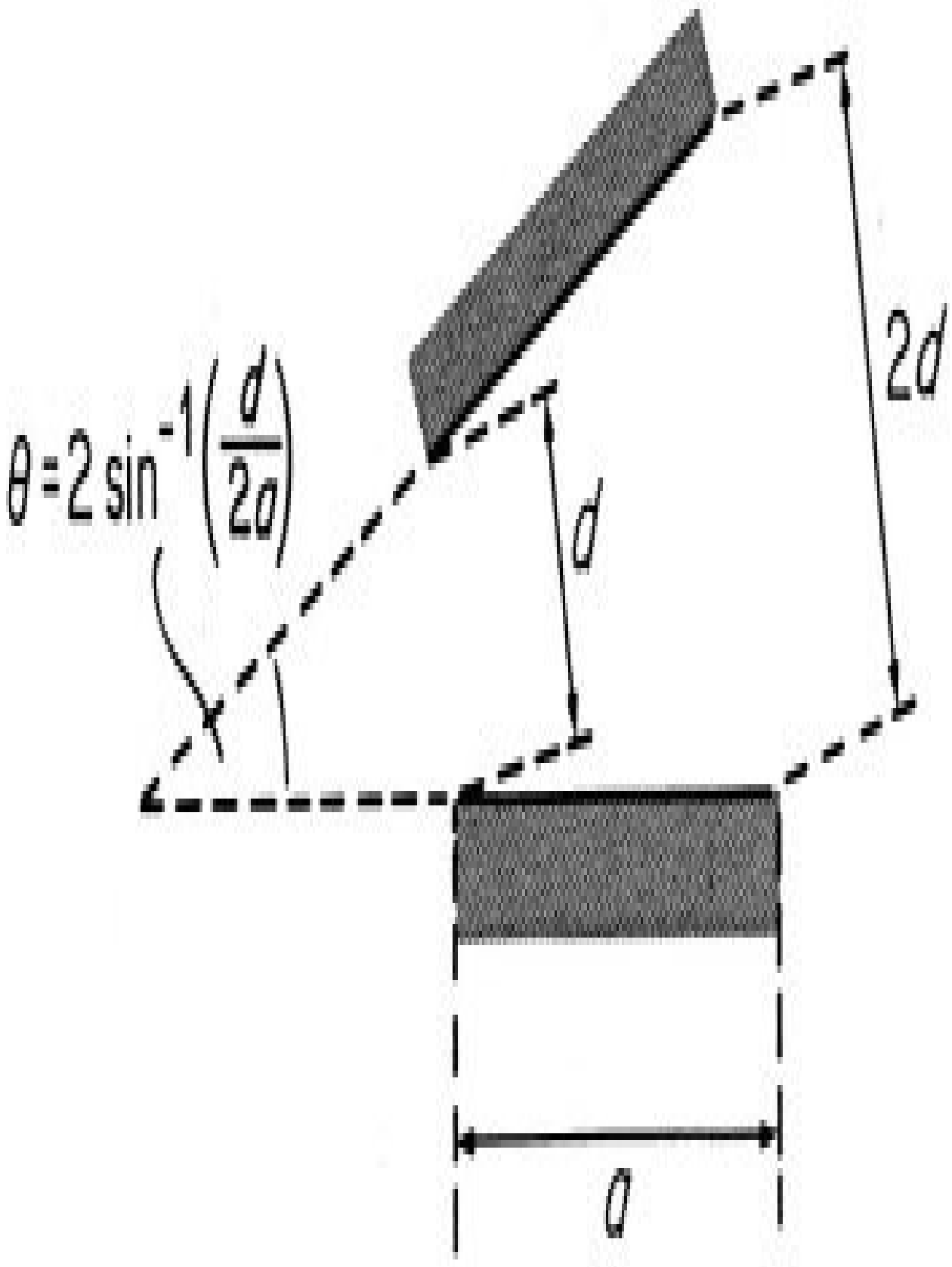
3.8 A hemispherical bulge protrudes from the otherwise flat plate of a parallel plate capacitor. Away from the bulge there is a uniform field in the z-direction. Choosing the origin as in Figure 3.14, show that the addition of a potential proportional to $\cos\theta/r^2$ leads to a solution of Poisson's equation in the neighbourhood of the bulge. Hence show that the field at the top of the bulge is three times as great as it is above a completely flat plate.

Figure 3.14.


z
r
θ

3.9 Two long rectangular plates of width a are arranged as shown in Figure 3.15 to form a wedge-shaped capacitor. Neglecting edge effects, the field lines between the plates are arcs of circles. Find an expression for the electric field and show that it satisfies Laplace's equation and the boundary conditions of the problem. What is the capacitance per unit length of the capacitor?

Figure 3.15.


$\theta = 2\sin^{-1}\left(\frac{a}{2a}\right)$
$2d$
d
a

3.10 (Estimate of capacitance from an approximate field pattern.) In the capacitor described in Problem 3.9, the field lines are actually arcs of circles. However, if the wedge is narrow, these arcs are not much different from straight lines. Assuming that the field lines are indeed straight, estimate the electric field as a function of position along the capacitor when there is a voltage V across the plates (neglect end effects). Then work out the total field energy stored by the capacitor and derive an approximate capacitance.

* E. R. Williams, J. E. Faller and H. Hill, Phys. Rev. Letters 26, 721 (1971)

* When deriving this potential the line charge is assumed to be infinitely long, and hence carrying an infinite charge. It is therefore not possible to choose the potential to be zero at infinity. The best that can be done is to take the potential as zero at some radius b, which may be as large as we please.

* The total charge carried by the pair of line charges is zero, and no logarithmic infinities arise if the potential is chosen to be zero at an infinite distance from the z-axis.

* Finite difference methods do not work near edges and corners. In the computed solution shown in Figure 3.12 the electrodes are thick, and the apertures are rounded. This of course corresponds more closely to reality than does the idealized thin aperture.

CHAPTER 4

Steady currents and magnetic fields

The previous chapters were concerned with the forces between stationary charges and their description in terms of the electrostatic field and potential. We now consider what happens when charges are in motion. As well as the force represented by the electric field, a moving charge may experience an additional force, called a magnetic force, when it is in the presence of other moving charges or permanent magnets. Magnetic forces are also conveniently described in terms of a vector field—the magnetic field—from which we can derive the force on a moving charge at any point.

A steady electric current in a conducting wire arises from the continuous motion of charge along the wire, and a moving charge near a current-carrying wire may therefore experience a magnetic force. In this chapter we shall first explain the rules governing the flow of current in conductors, and then go on to discuss the magnetic fields caused by steady currents in conductors.

4.1 ELECTROMOTIVE FORCE AND CONDUCTION

4.1.1 Current and resistance

Some of the electrons inside conductors are not bound to particular atoms but are free to move from atom to atom. In a metal we may picture these free electrons, the conduction electrons, as behaving more or less like gas molecules in a box, moving about at random with large speeds, and colliding with atoms in the metallic lattice. In the absence of an applied electric field the mean electron velocity, averaged over any volume which is very large on a microscopic scale, will be zero. If an electric field E is maintained in the conductor the electrons acquire an average velocity v in the direction opposite to the field. This is because, between collisions with atoms in the conductor, the electrons are acted upon by a force – eE. The mean drift velocity v of the electrons is thus equal to half of the average extra velocity given to electrons in between collisions.

A conductor carrying current can be, and usually is, electrically neutral. The number of free electrons is balanced by an equal number of positively charged atoms. These atoms are also subjected to an electric force, but they are unable to move, being held in their lattice sites by far stronger interatomic forces. In liquid and gaseous conductors there may exist moving carriers with charge of either sign. We restrict the discussion here to solid conductors.

The electron flow in a solid conductor can be described at every point by the current density, j. The current density is a vector function of position; at any point the electrons flow in the direction of their mean drift velocity at that point. Because the charge on the electron is by convention negative, the direction of the vector j is opposite to the direction of the velocity v. The magnitude of the current density is the net amount of charge per unit time crossing unit area of a surface perpendicular to the mean drift velocity. Hence, if the number of free

electrons per unit volume in a conductor is N, the current density is given by

(4.1)

The current I through any surface S is the net charge that passes through the surface per second. It is given by the surface integral of the current density

(4.2)

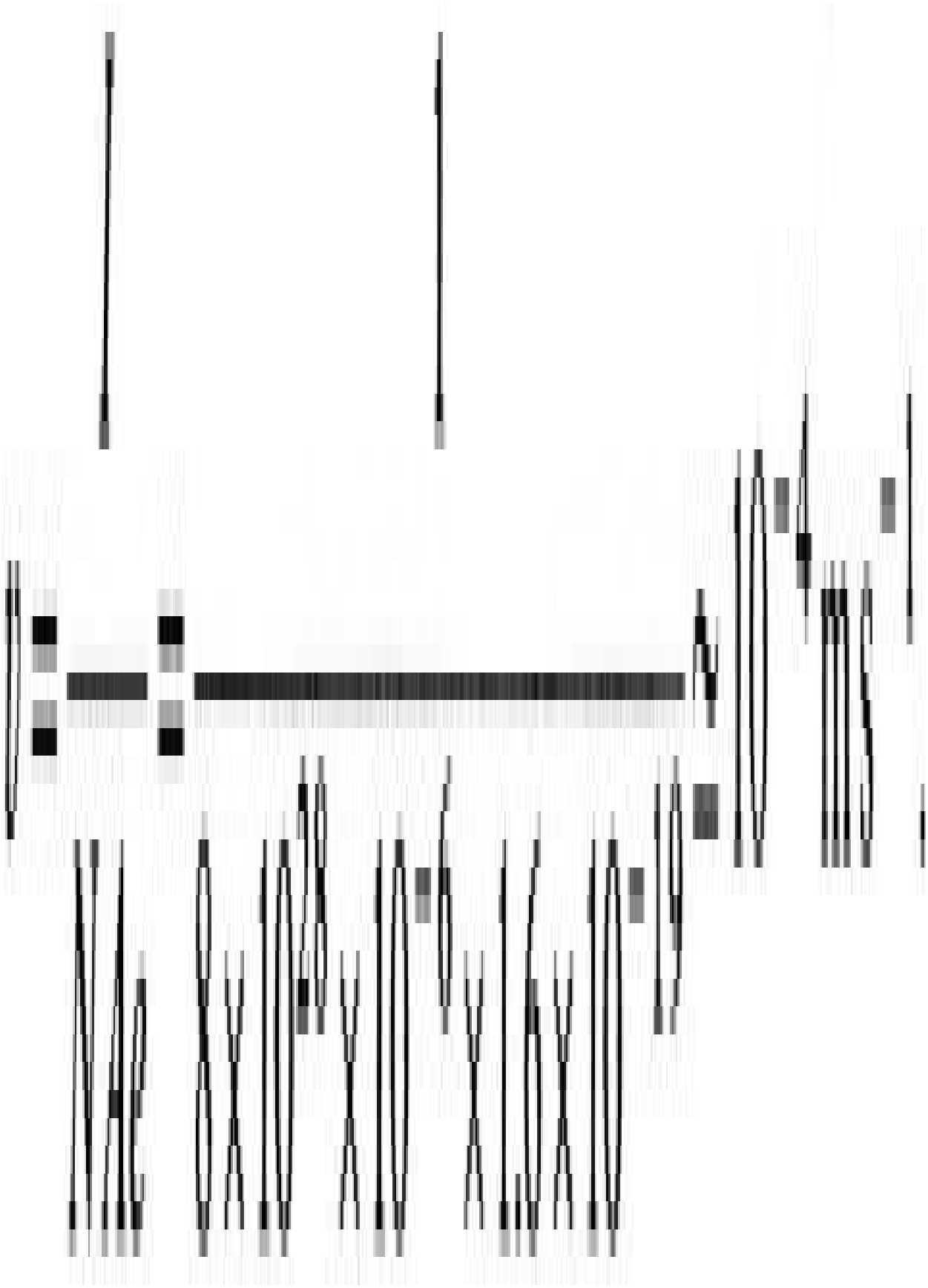
In a conductor carrying a steady current the surface integral has the same value for any surface that cuts across the flow of charge. For example, when a wire joins the two terminals of a battery, the current flowing into the wire at one terminal equals the current flowing out at the other terminal. The current in a conductor is thus the total amount of charge that enters or leaves the conductor per second. Current has the dimensions [charge] [time]−1, and is measured in amperes (A). Current density has the dimensions [charge] [time]−1 [area]−1, and is measured in A m−2.

Let us estimate the mean drift speed of the electrons in a typical conductor carrying a current. If a current I flows in a metal wire of uniform cross-sectional area A the current density j will be constant and parallel to the length of the wire. The current I is thus given by Equation (4.2)

(4.3)

In a good conductor there is usually about one free electron per atom; hence in a

copper wire for example we obtain N $8 \times 10^{28} m^{-3}$. The electronic charge e is equal to $1.6 \times 10^{-19} C$, and so for a wire of cross-sectional area $1 mm^2$ carrying a current of 1 A,



This calculation illustrates that the drift velocities of electrons in a solid are very small. The random thermal velocities upon which the mean drift velocity is superposed are always very much larger.

For many materials it is found that in the steady state the current density is proportional to the electric field. For such materials

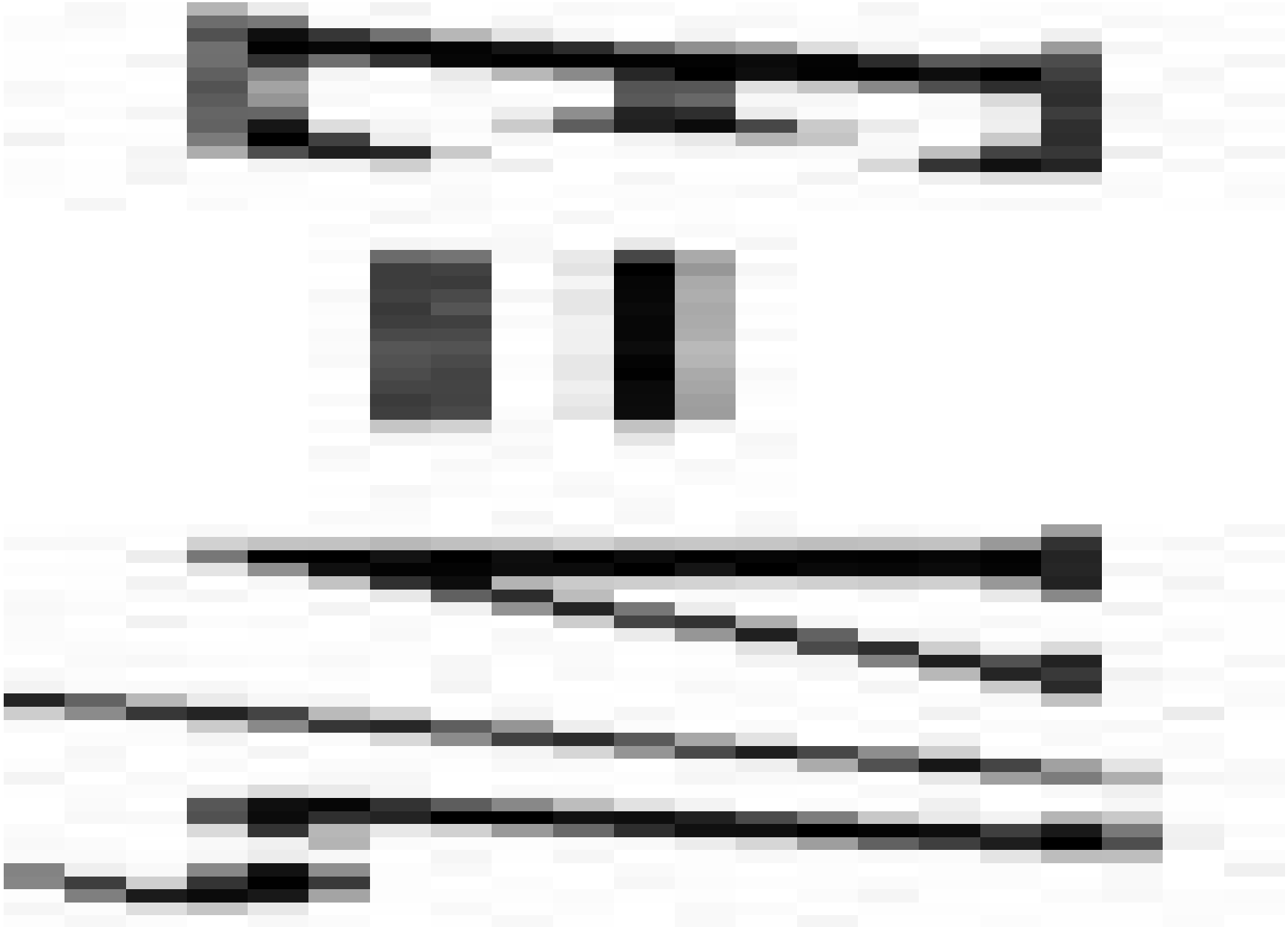
(4.4)

where σ is constant for the material at a given temperature, and is called its electrical conductivity. Materials that obey Equation (4.4) are said to be ohmic, and Equation (4.4) is called Ohm’s law.

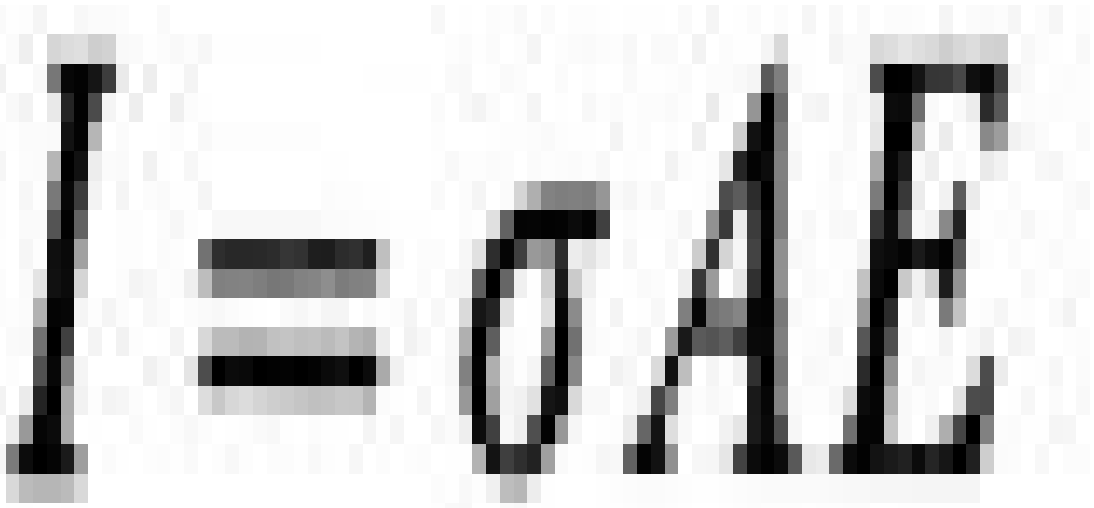
Consider a homogeneous conductor of uniform cross-sectional area A in which a steady current I is flowing. For all surfaces S cutting the wire

(4.5)

If a potential difference of V volts is maintained across the ends of the conductor by an external source, the electric field E will be uniform and



where l is the length of the conductor. Equation (4.5) now becomes



The ratio of the voltage to the current in the wire

(4.6)

(4.7)

Table 4.1 The conductivities, of some common materials at 293K.

MATERIAL	CONDUCTIVITY (ohm m)−1
Copper	5.9×10^7
Gold	4.1×10^7
Germanium (pure)	2.2
Saturated NaCl solution	23.0
Glass	10−10−10−14
Fused quartz	1.3 × 10−18
Wood	10−8 −10−11

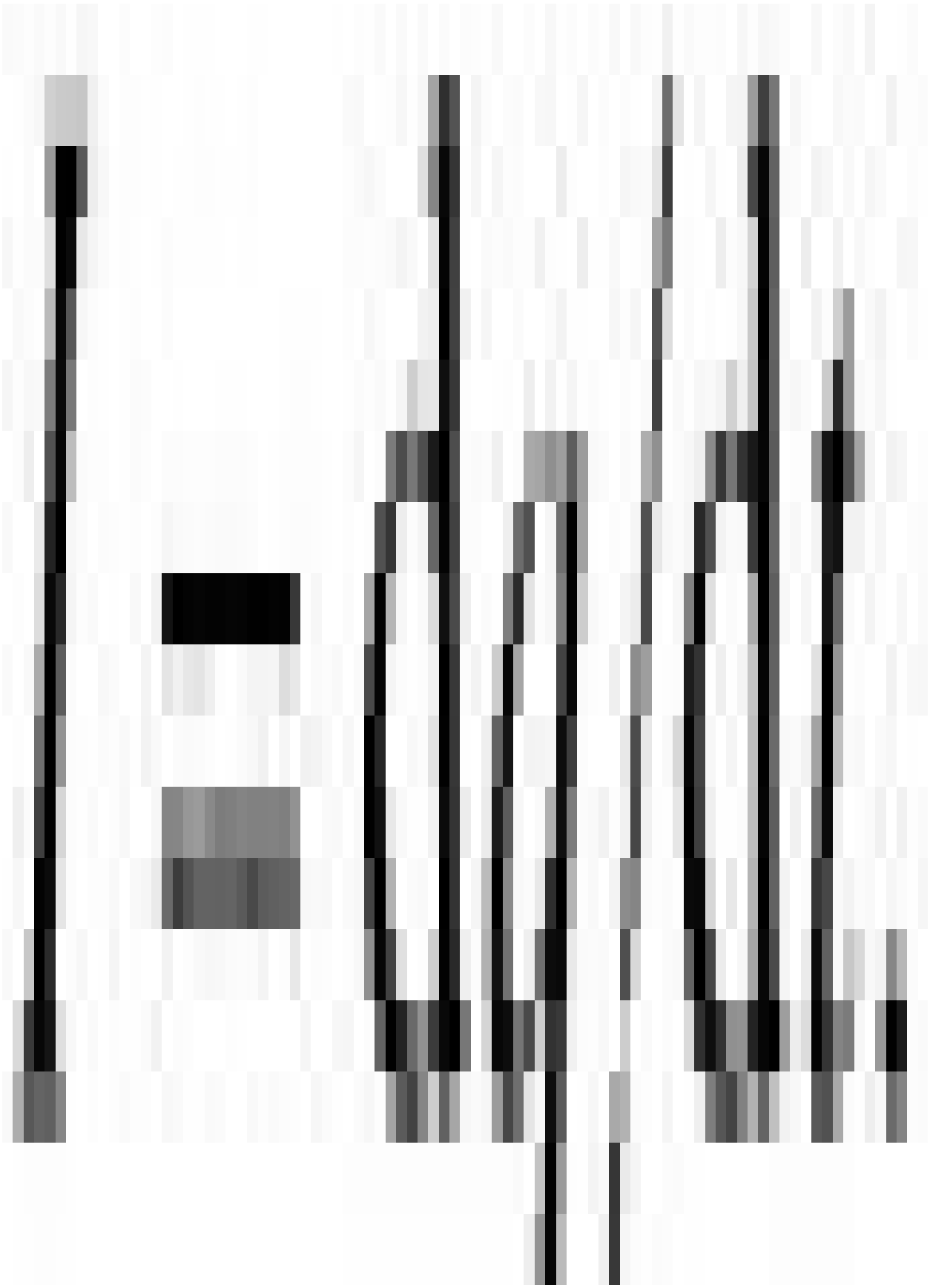
where R is called the resistance of the conductor. R is measured in ohms (symbol Ω). The conductivity σ is measured in (Ω m)−1.

In the more general case of a conductor of arbitrary shape, the electric field inside is not constant when a current I is flowing. However, as long as Equation (4.4) is obeyed inside the material the current I remains proportional to the voltage V across the ends of the conductor, and the resistance is still defined by Equation (4.7). Ohm's law is not always valid. Gases obey Ohm's law at low electric fields, but not at high fields; some materials at low temperatures deviate from Ohm's law to the extent that their electrical conductivities become infinite. Even for ohmic materials the conductivity of a material is usually not a constant but is very dependent on temperature, and this is especially true for semiconductors like silicon and germanium. Pure germanium has a conductivity of 2.5 (Ωm)−1 at 300 K, but at 1000 K the conductivity has risen to 3×10^4 (Ωm)−1. Table 4.1 gives the electrical conductivities of several comon materials at 293 K, and illustrates the enormous range over which the values of the conductivities spread. The reciprocal of the conductivity of a material is called its resistivity. The resistivity is denoted by ρ, and is measured in ohm metres. We adopt the commonly used symbols ρ and σ for the resistivity and the conductivity, although we have already used them for volume and surface charge densities. It is always clear from the context which meaning the symbols have.

We have seen in Chapter 1 that if charges are placed on an isolated conductor the electric field inside becomes zero, and the charge is distributed over the surface of the conductor. For a steady current to flow in a conductor an external agency must maintain an electric field inside the conductor. The agency that does this is called a source of electromotive force (e.m.f.). A common source of steady e.m.f. is a battery. A battery of e.m.f V volts maintains a potential difference of V volts between the ends of a length of conductor connected across the battery terminals. In this situation a steady current flows round the complete circuit formed by the conductor and the source.

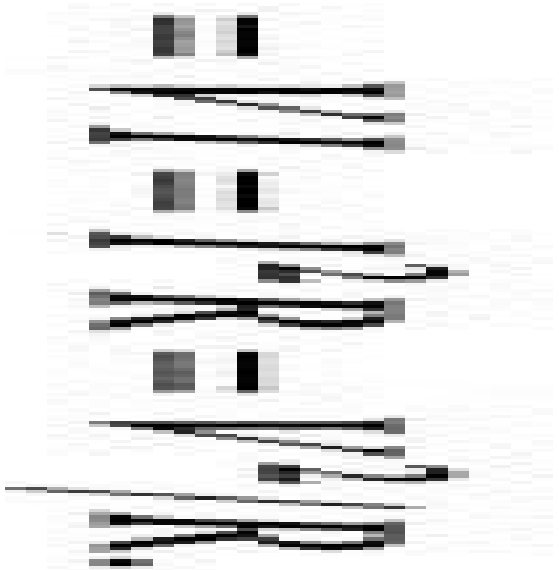
There are several types of sources of steady e.m.f.s and of e.m.f.s which change with time. We shall not explain the principles of operation of different sources of e.m.f., but simply point out that they all convert some other kind of energy into electrical energy. A battery, for example, uses stored chemical energy to drive a current around a circuit. The chemical energy reappears as heat dissipated in the resistors. This heating arises because part of the energy the electrons acquire by virtue of the work done on them by the applied electric field is given up to the atoms with which they collide.

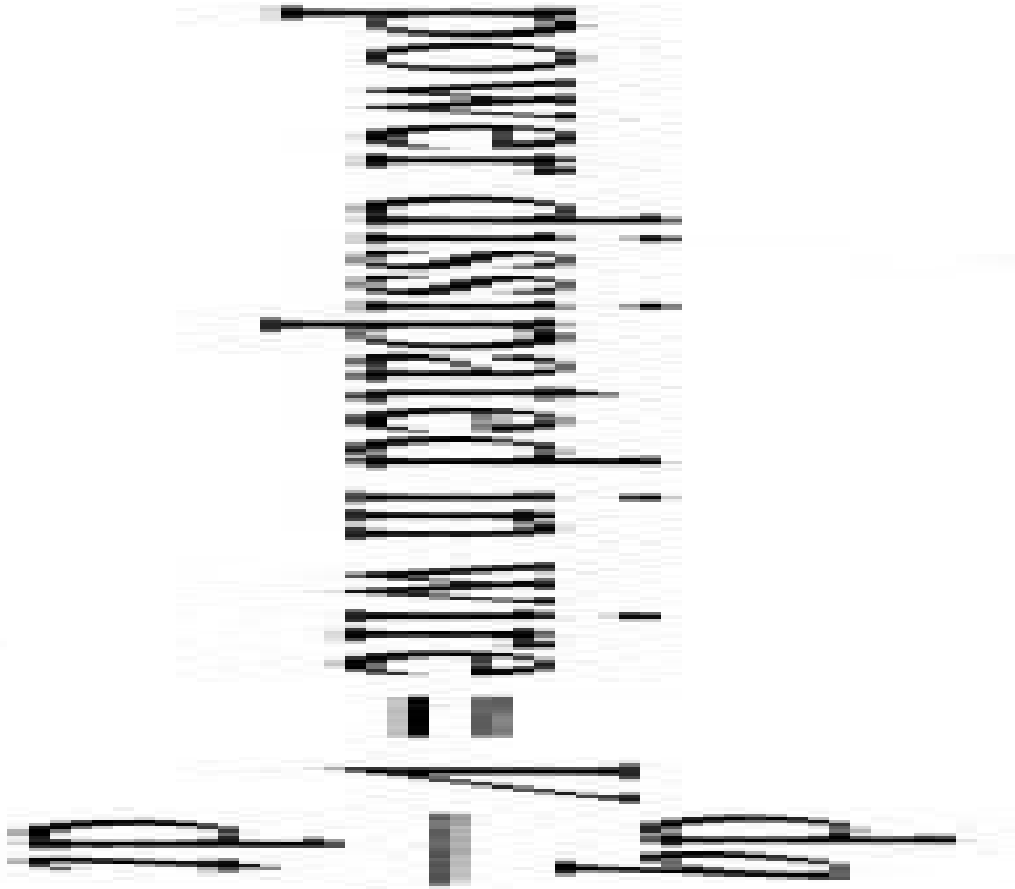
The power dissipated in a resistor can easily be worked out. Suppose that a steady current I flows in a wire which has a potential difference V between its ends. If, in time dt, a charge dq passes through any surface that cuts through the wire, then



In the steady state a charge dq leaves the wire at one end and the same amount of charge enters it at the other end. The net effect is that during the time dt a charge dq has lost an amount V dq of potential energy, and

(4.8)





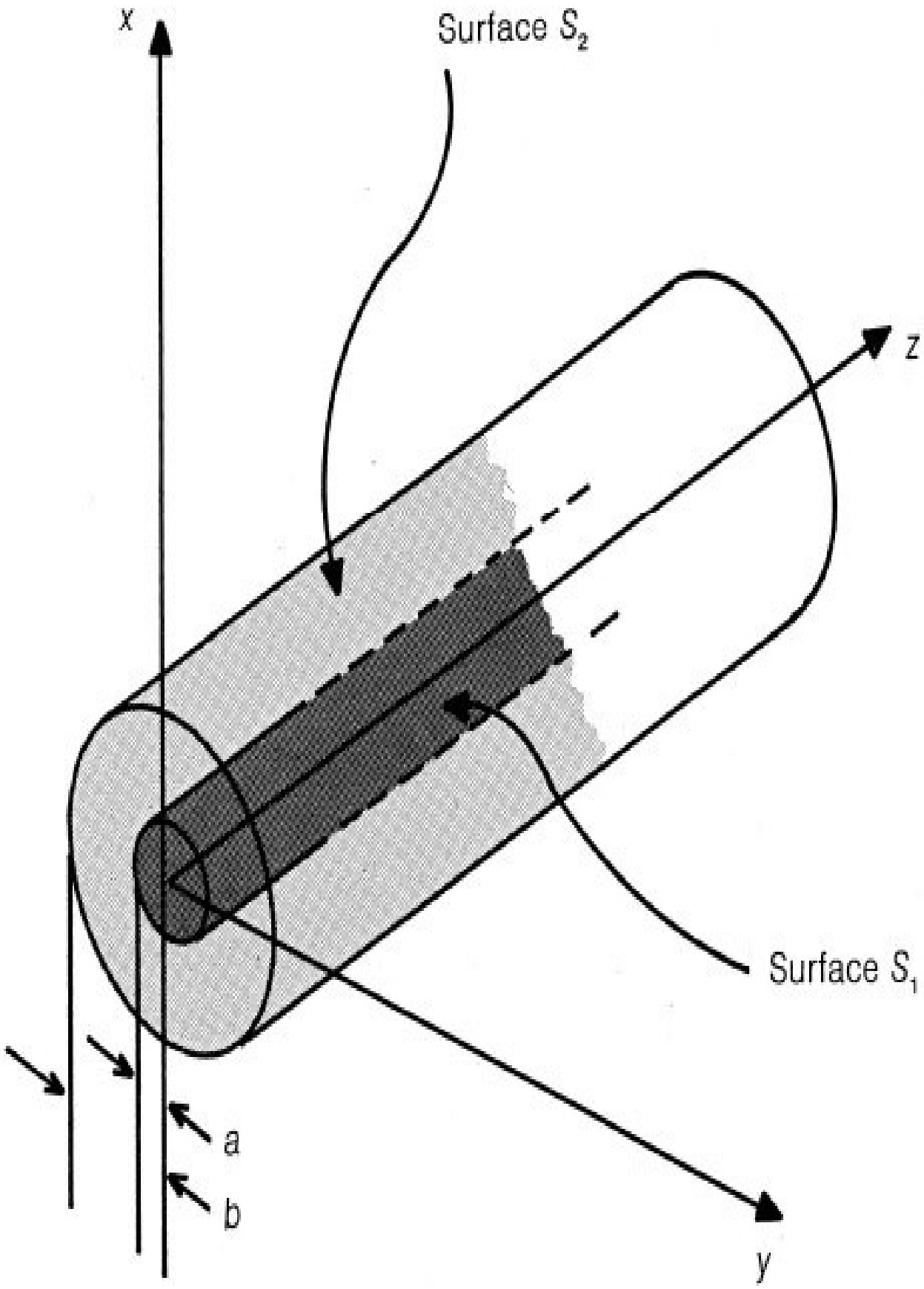
Before closing this section we make a cautionary remark about the above discussion of currents in conductors. The picture of conduction electrons moving freely in conductors is an oversimplification, although it gives a good qualitative description of many of the properties of electric currents. In some materials the current is due to the motion of carriers with a charge equal in size to that on the electron but of opposite sign. These carriers are called holes and can be regarded as gaps or holes in a conductor which otherwise has a full complement of electrons. They have positive charge because a conductor with a full complement of electrons would be electrically neutral due to neutralization of the total charge on the electrons by the positive charge on the atoms from which the electrons came. The proper description of metallic conduction requires the use of quantum mechanics, and we will not consider it in this book.

4.1.2 The calculation of resistance

The calculation of the resistance between two electrodes on a piece of conducting material of arbitrary shape is not always easy, even if the material is uniform and obeys Ohm's law. We will outline the technique used in the general case by considering a simple example.

We will calculate the resistance between the inner and outer conductors of the cable shown in Figure 4.1. The inner wire is of radius a and lies along the axis of the outer conducting sheath of radius b, which is earthed. The space between

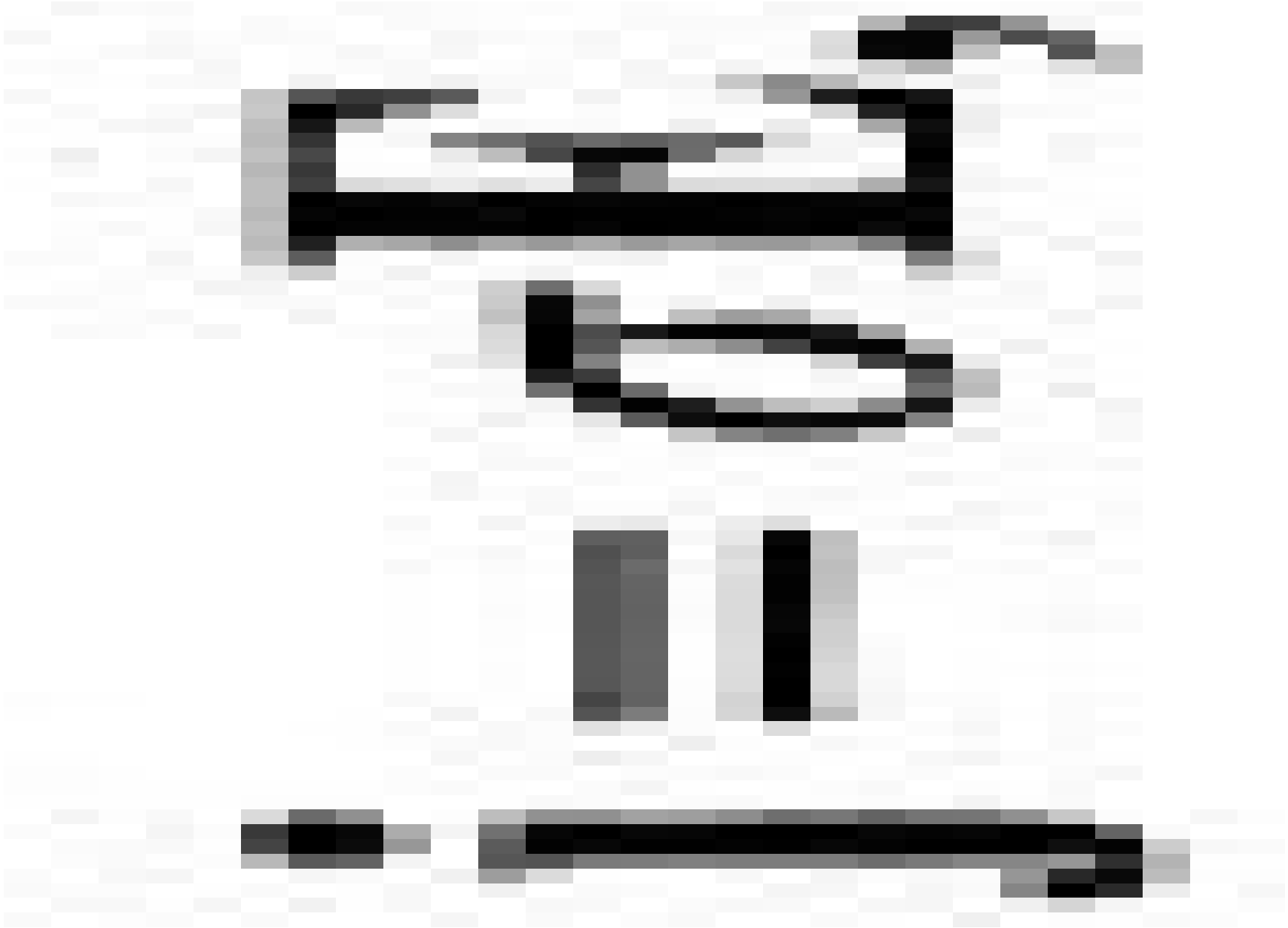
Figure 4.1. A cable consisting of coaxial cylindrical conductors. The space in between them is filled with an imperfect insulator.


x
Surface S_2
z
Surface S_1
a
b
y

the conductors is filled with a material which is not a perfect insulator, but has a conductivity σ. What is the resistance between the inner and outer conductors of a length L of the cable? This resistance may be of practical significance since it determines the leakage current to ground when the cable is used to transmit a voltage along the inner wire. To determine this resistance we have to calculate the current I that flows from the inner wire to the outer sheath, over a length L of the cable, when the inner wire is at a potential V. Now

(4.9)

where j is the current density in the insulator at the surface S1 of a length L of the inner wire. (The current could be obtained by taking the surface integral of the current density over any surface between the conductors that encloses the inner wire. We choose surface S1 merely for the sake of convenience.) From Equation (4.4)



where E is the electric field over the surface S1. This field is perpendicular to the metal surface and by symmetry has a constant magnitude E(a). Equation (4.9) can now be written

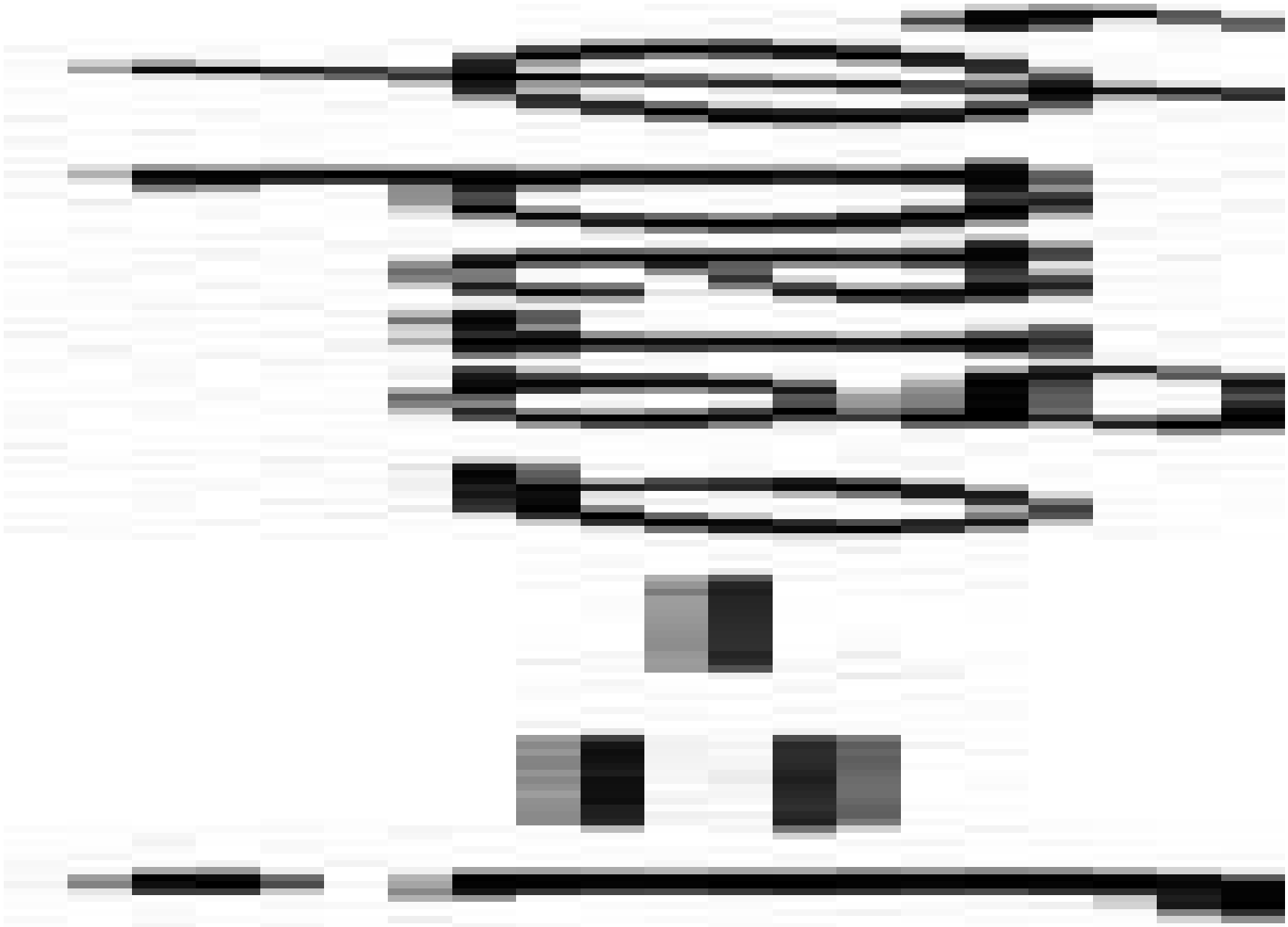
(4.10)

and the problem reduces to the determination of the field E in terms of the quantities V, a and b.

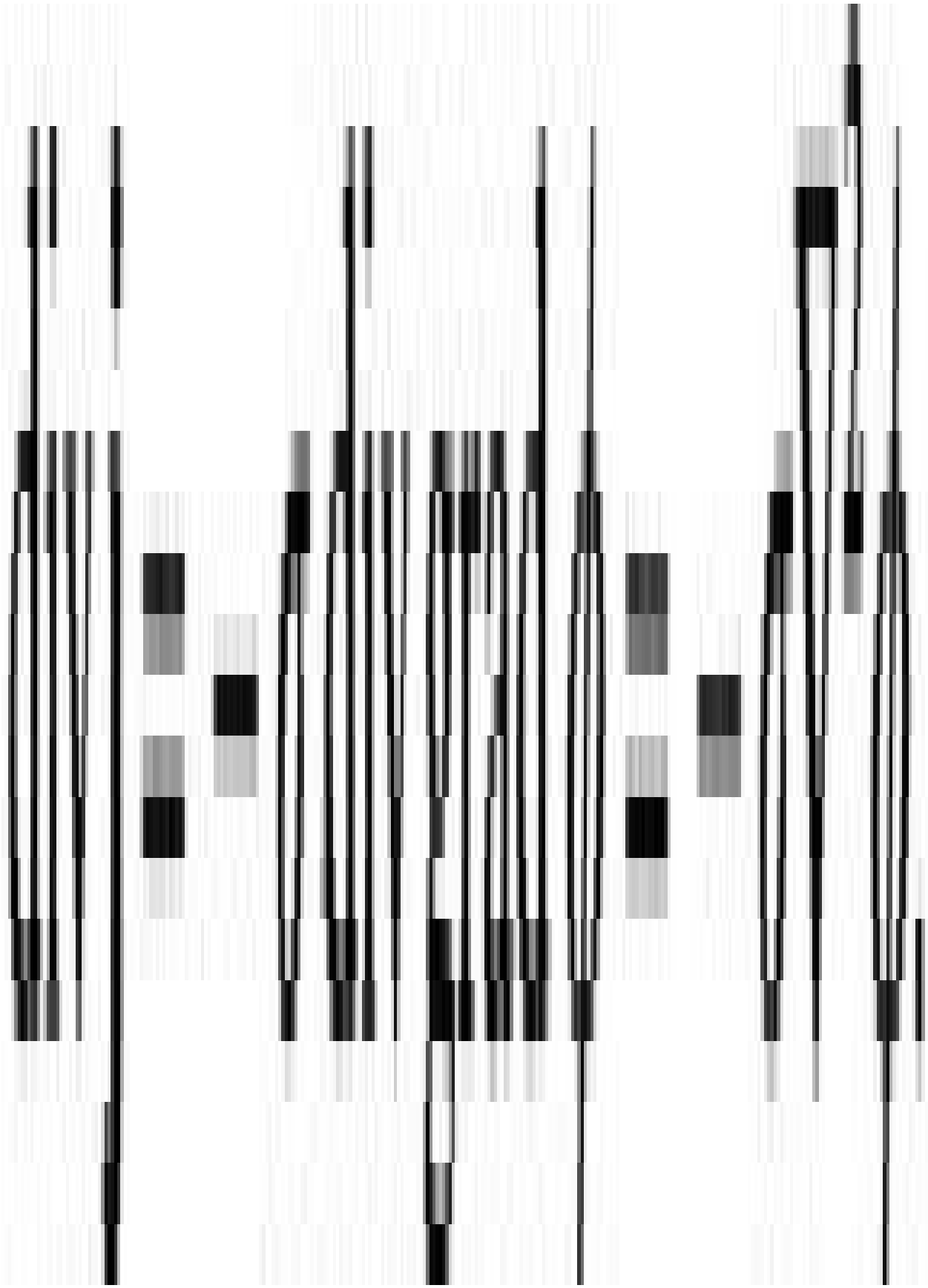
When a source of e.m.f causes a steady current to flow, the electric field outside the source is everywhere static. The field E can thus be written in terms of the electrostatic potential ϕ according to Equation (1.26),

(4.11)

We now need an equation from which we can determine the potential everywhere within the material Combining Equations (4.4) and (4.11) gives



or



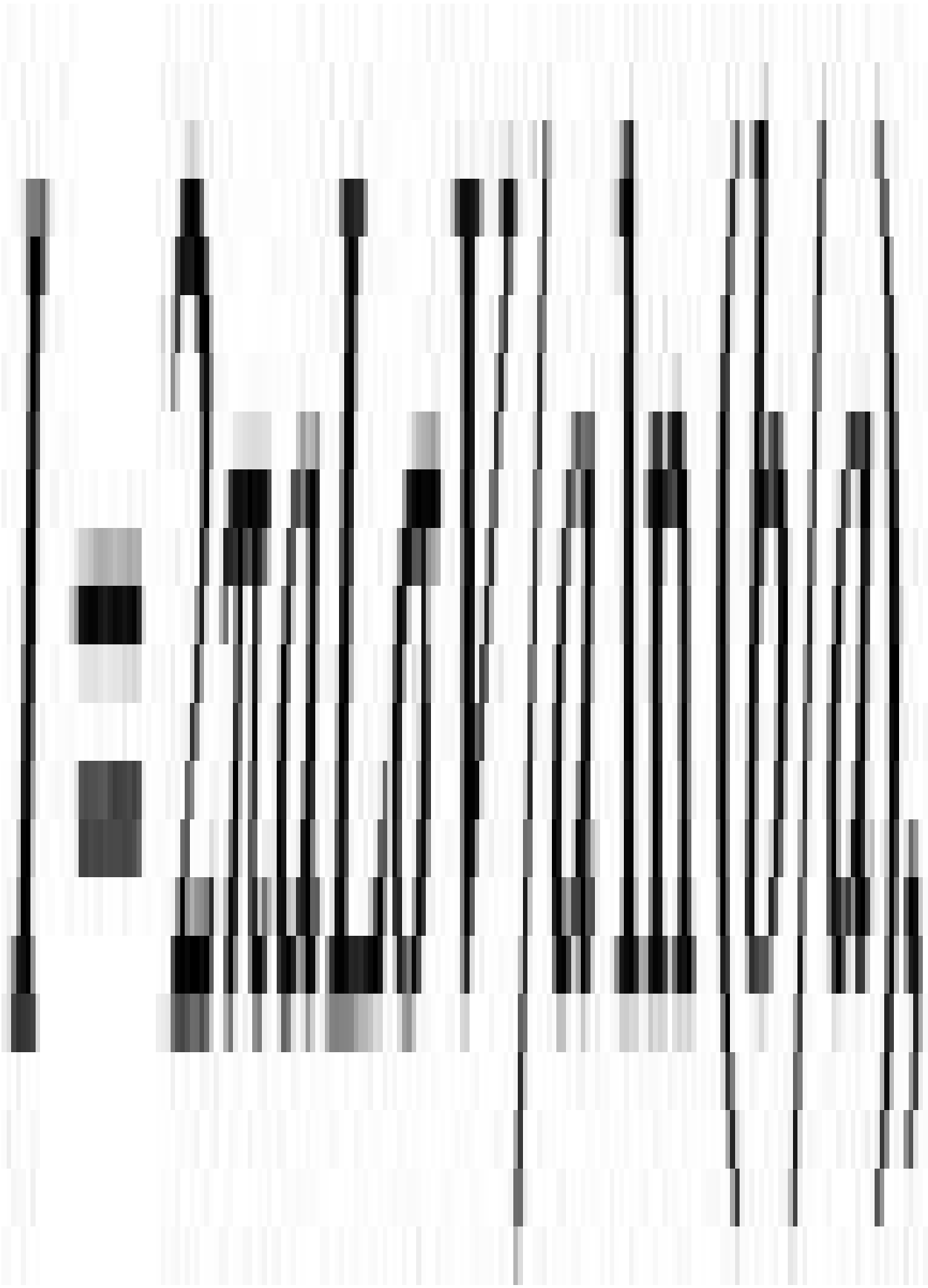
The divergence of the current density at any point is a measure of the flux of the vector j that flows out over the surface of a very small volume surrounding the point. If the current is steady as much charge enters any small volume within the material as leaves it, and the divergence of j is everywhere zero. Hence the equation

(4.12)

is obeyed everywhere within the conducting medium. Equation (4.12) is Laplace's equation, the same equation for the electrostatic potential as we derived in regions in which no current flows and there is no volume density of free charge. In the present problem we require a solution to Equation (4.12) which obeys the boundary conditions that the potential has the constant value V at all points on the surface S1, and the constant value zero at all points on the surface S2. This problem has already been solved in section 3.1. For a very long cable the potential ϕ does not vary in the z-direction. Since the cable is cylindrically symmetrical ϕ is independent of the azimuthal angle θ, and depends only on the distance r from the z-axis. The required solution of Laplace's equation is

(4.13)

where A and B are constants. The solution has been determined in section 3.2 and is given by Equation (3.5). Using this successively in Equations (4.11) and we find



and hences

(4.14)

A typical insulator has a conductivity of about $10^{-12}(\Omega m)^{-1}$. The leakage resistance of 100 m of a cable whose inner wire has a radius of 1.5 mm and outer wire a radius of 10 mm, if filled with such an insulator, would be about $3 \times 10^{9}\Omega$.

This example of the calculation of resistance illustrates the analogy between the solution of steady current problems and the solution of electrostatic problems in free space. In each case a solution of Laplace's equation is required, with the boundary conditions that the potentials over metallic surfaces are constant. In the calculation of resistance the metallic surfaces are the electrodes through which the current enters and leaves the resistive material. In the analogous electrostatic problem the metallic surfaces act as the two conductors of a capacitor.

We emphasize that the calculation of resistance made above is valid only when Ohm's law is obeyed by the material. The steps between Equations (4.11) and (4.12) require that Ohm's law holds, and this is not always so. For example, the law breaks down for many materials at high electric fields.

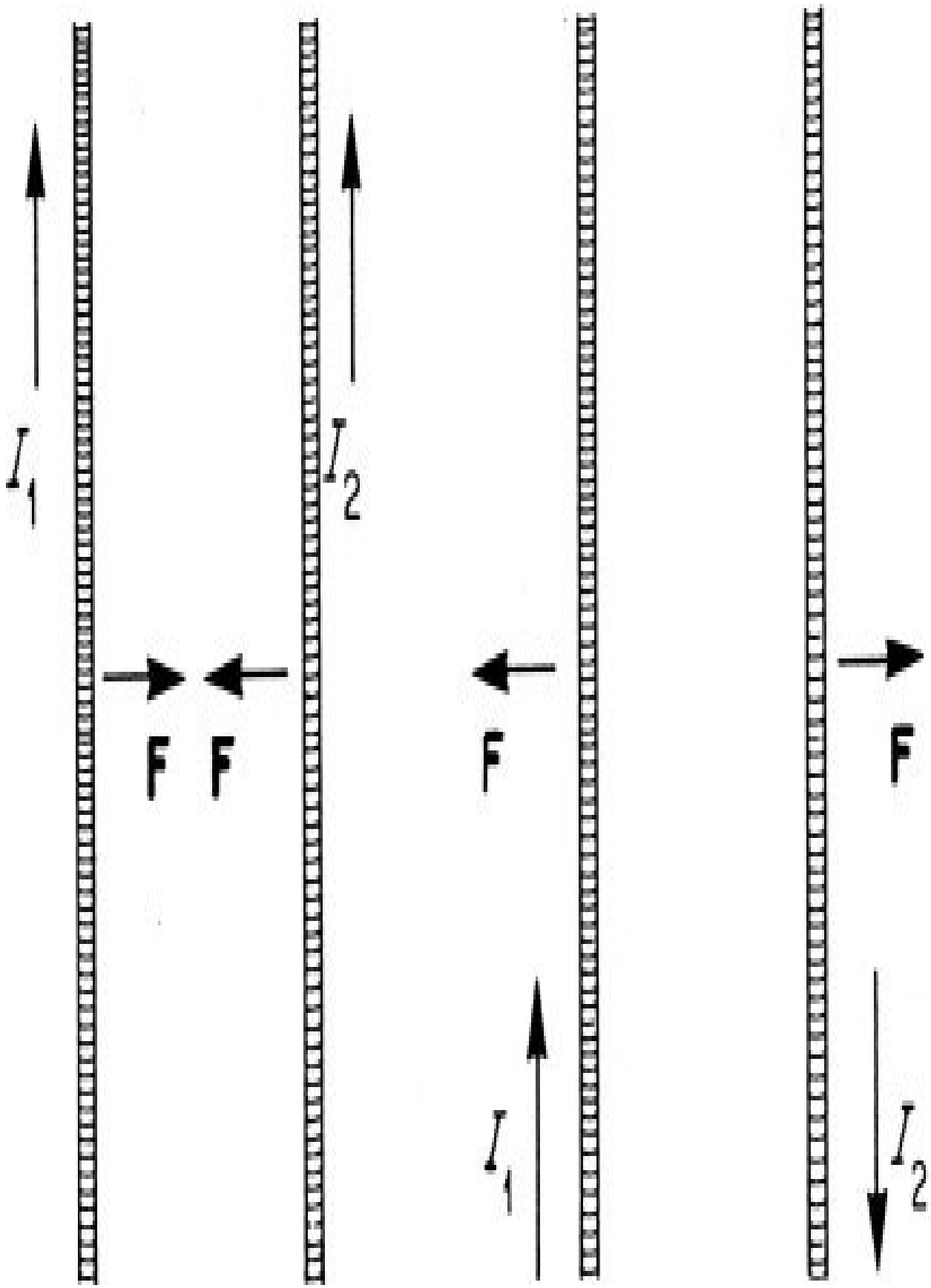
4.2 THE MAGNETIC FIELD

4.2.1 The Lorentz force

As well as the forces exerted on charges by electric fields, there may be additional magnetic forces experienced by moving charged particles. Magnetic forces always act perpendicularly to the direction of motion of the charged particles. For example, if two parallel metal wires are brought near to each other and steady currents are set up in them, there is a force between the two as shown in Figure 4.2. If the currents are in the same direction the force is attractive; if the currents are in opposite directions the force is repulsive. The force between the wires is not the same as the electrostatic force since the wires contain equal amounts of positive and negative charge and are electrically neutral.

Permanent magnets also cause forces to act on wires carrying currents. If a current is passed through a wire suspended between the poles of a permanent

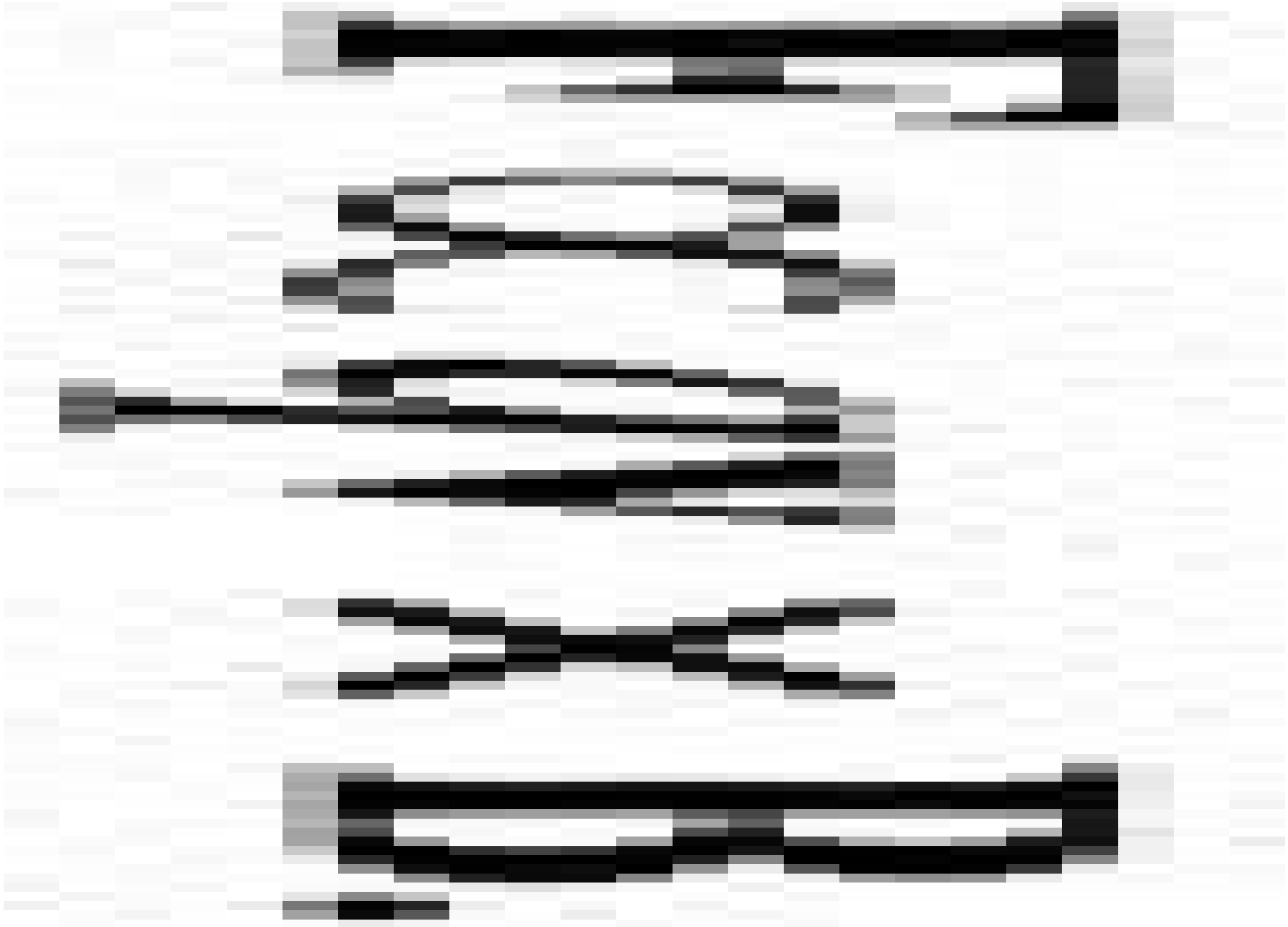
Figure 4.2 The forces between parallel wires carrying currents. Currents in the same direction cause attraction; currents in opposite directions cause repulsion.


I_1
I_2
F F
F
F
I_1
I_2

magnet the wire experiences a force perpendicular to the direction of the current and is deflected. If the direction of the current is reversed the deflection of the wire is reversed. A wire with a steady current flowing in it and a permanent magnet both have the same effect on another wire carrying a current. In order to describe these effects we introduce the magnetic field B (sometimes called the magnetic induction field).

The current I1 in the wire shown on Figure 4.2 gives rise to a field B 1 which describes the forces on the moving charges constituting the current I 2 in the other wire. The current I 2 can be seen from Equation (4.3) to be proportional to the product of the mean speed of the electrons and the electronic charge. Experiments show that the force on the wire carrying the current I 2 is proportional to I 2, for a fixed current in the first wire. Hence, in a constant magnetic field, the magnitude of the force on a moving charge is proportional to the product of the charge q and the speed υ of the charge.

We may do more experiments to determine how the force on a moving charge depends on its velocity and the magnetic field in which it moves. Let us make the reasonable assumption that the field B in the middle of the gap between the parallel pole faces of a large permanent magnet is uniform and in a direction perpendicular to the pole faces, as shown in Figure 4.3. Experiments with charged particles moving in the gap show that the particles always experience a force F which is perpendicular both to the field direction and to the velocity v of the particles. For example the electron beam of a cathode ray tube placed in the gap is deflected parallel to the pole faces. The magnitude of the force is found to be proportional to the speed of the particle, its charge q, and to the sine of the angle between the vectors v and B. These experiments are illustrated schematically in Figure 4.4. Their results can be summarized by writing



The units of the magnetic field B are defined by choosing the constant of

Figure 4.3. A uniform magnetic field between the pole faces of a permanent magnet. We represent the magnetic field in strength and direction with field lines, just as for the electrostatic field.

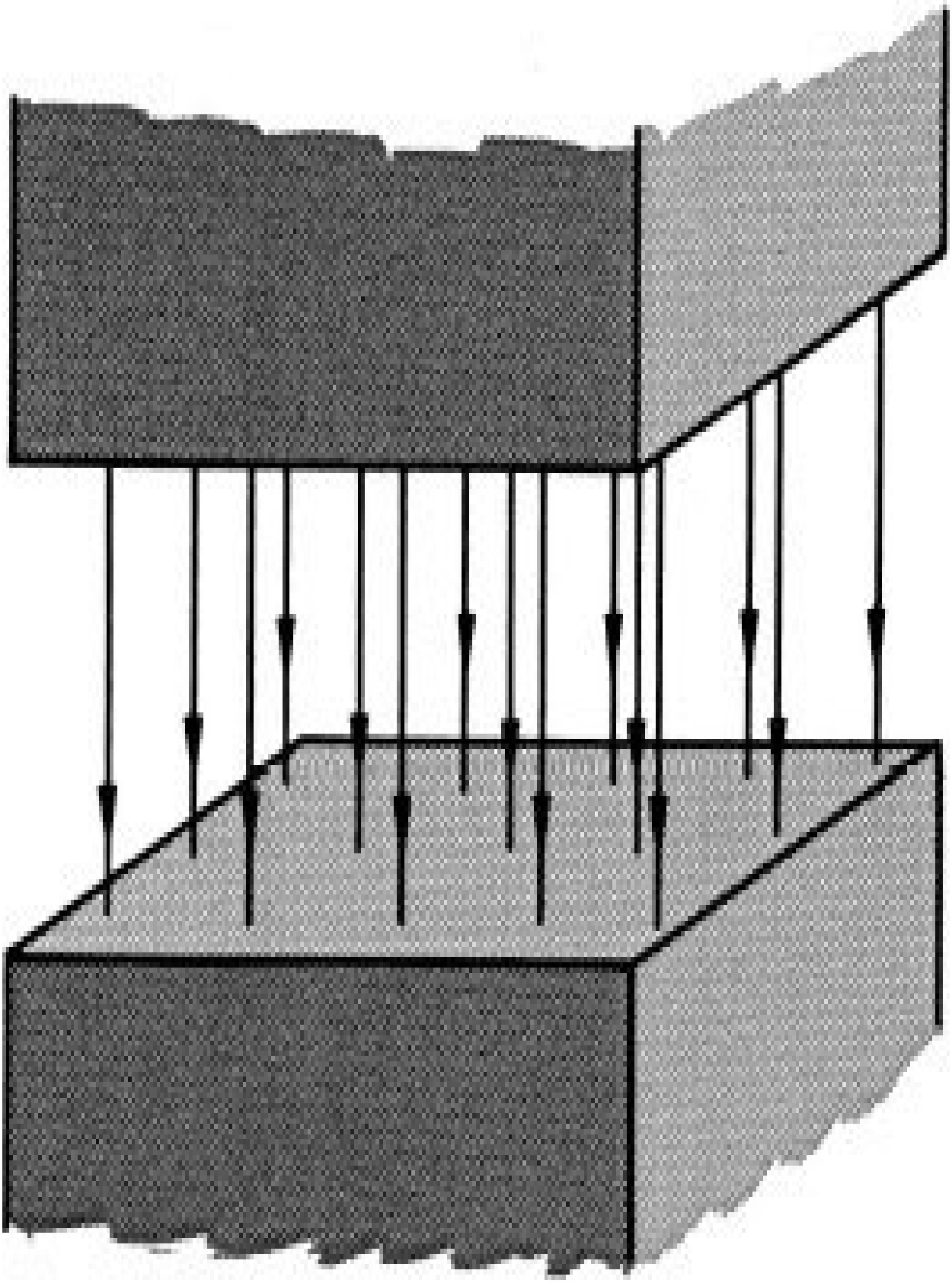
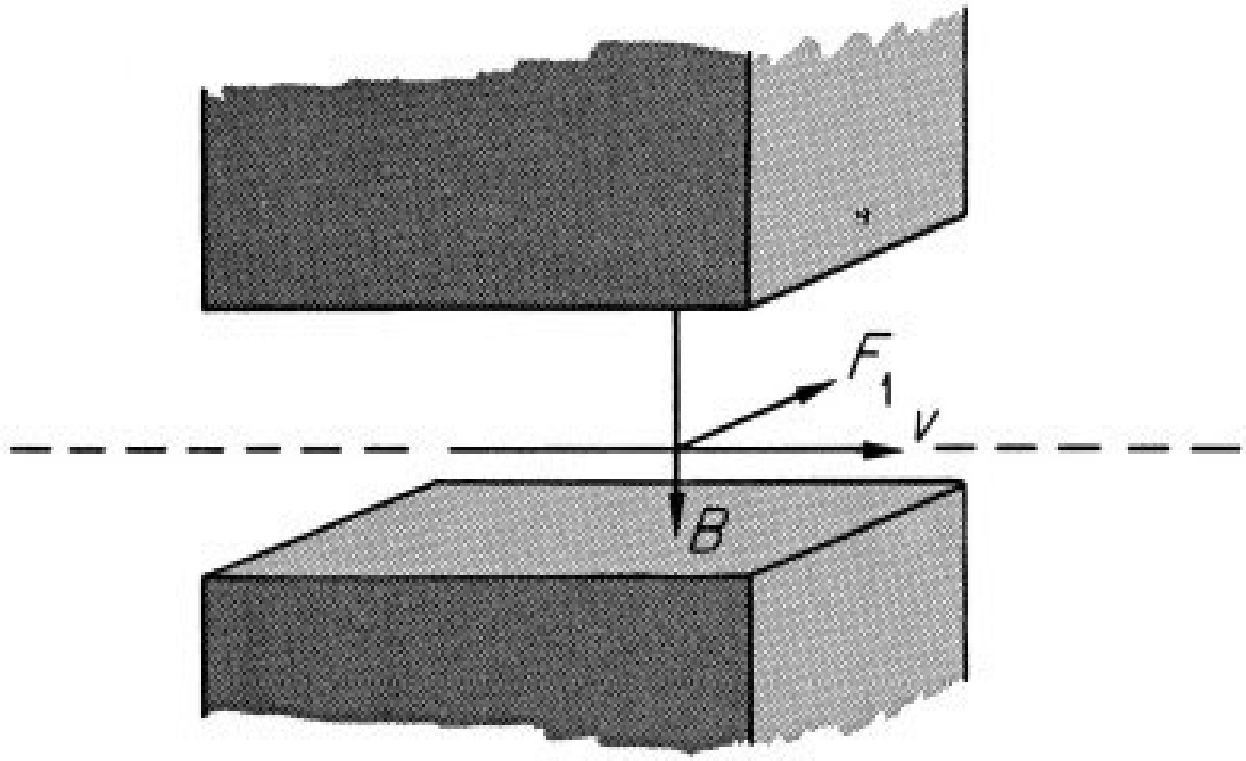
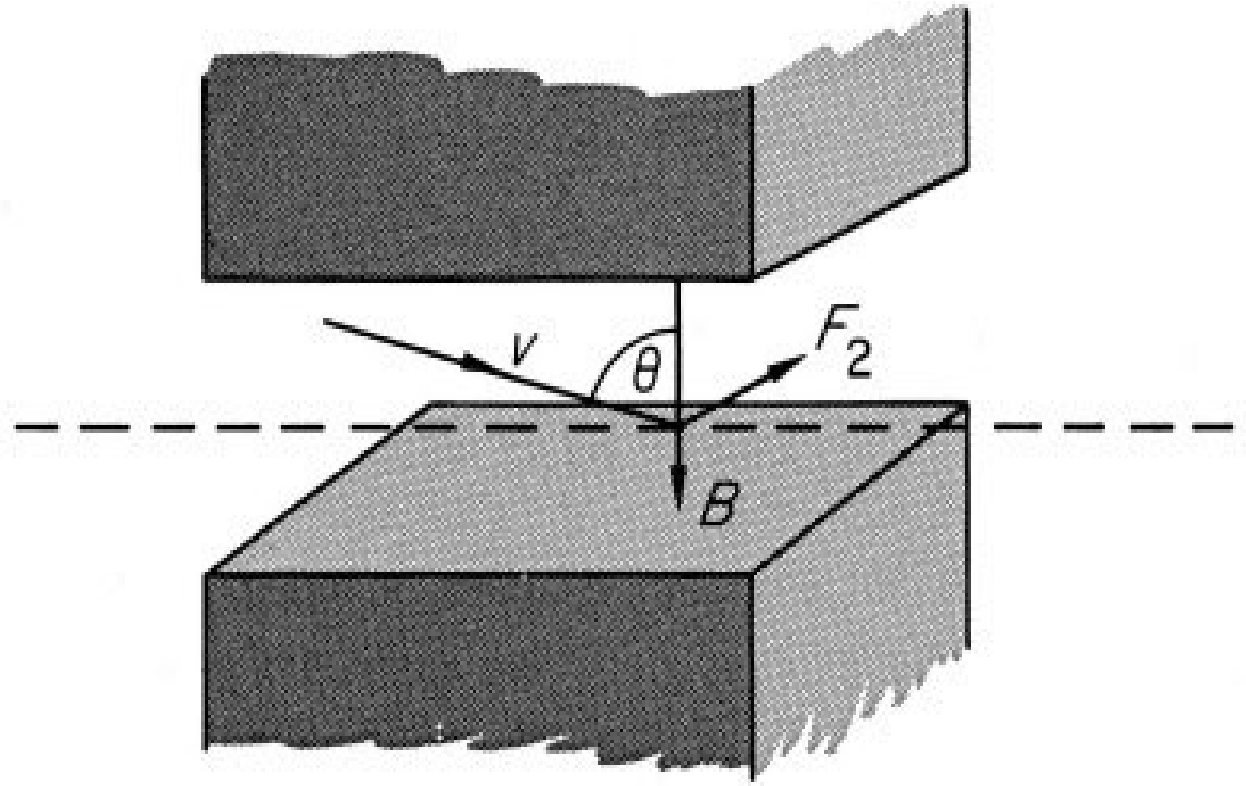


Figure 4.4. The force on a positively charged particle moving in a uniform magnetic field. If the particle moves parallel to the pole faces, as shown in the top diagram, the magnitude of the force F1 is proportional to the product υB. When the velocity v of the particle makes an angle θ with the direction of the field B the magnitude of the force F2 is proportional to υB sin θ.


F_1
v
B


v
θ
F_2
B

proportionality to be unity, leading to

(4.15)

This equation is known as the Lorentz force law. The direction of the force is illustrated in Figure 4.5. In a Cartesian coordinate system the force F has components:

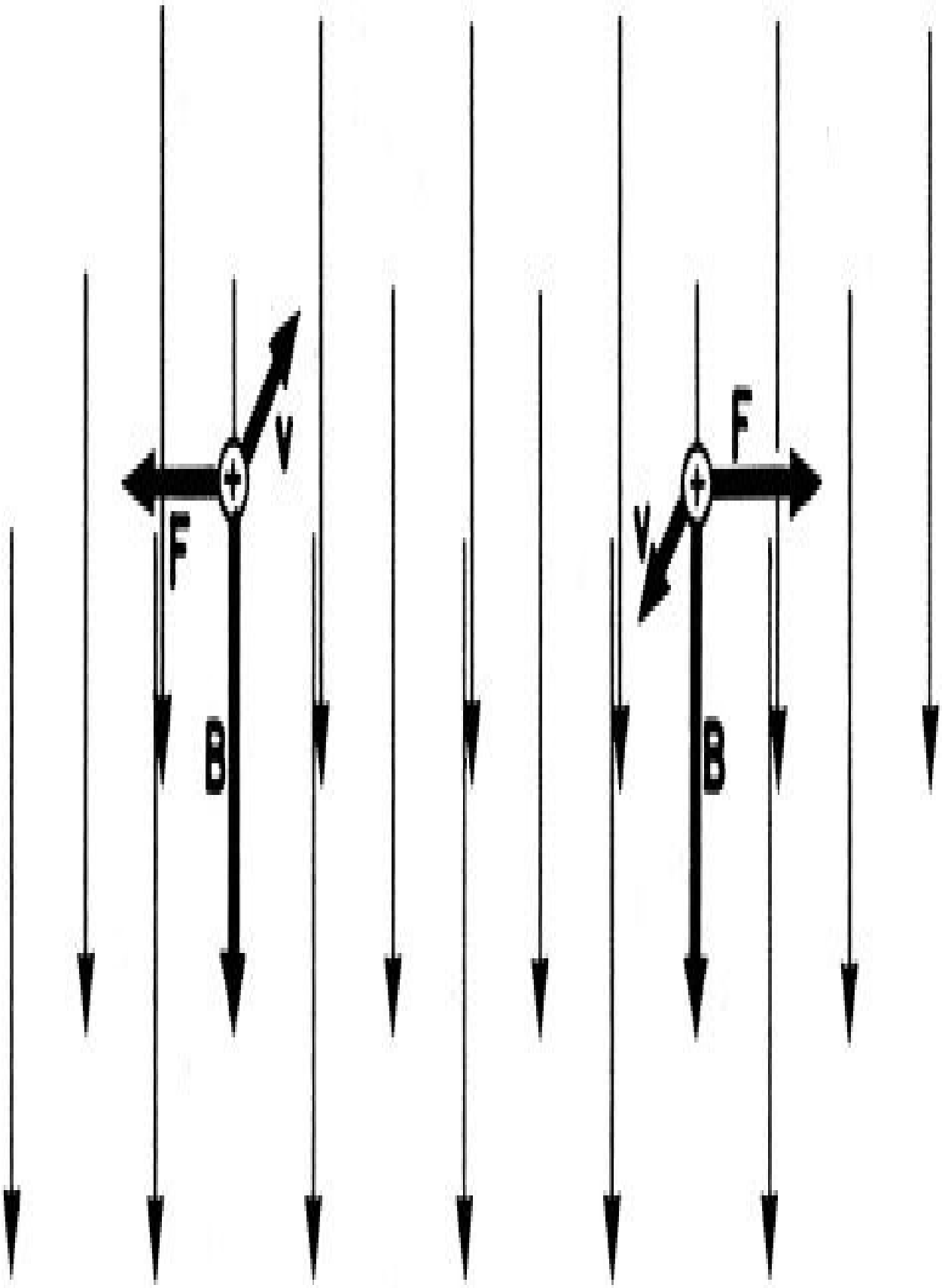
$$F_x = q(v_y B_z - v_z B_y)$$

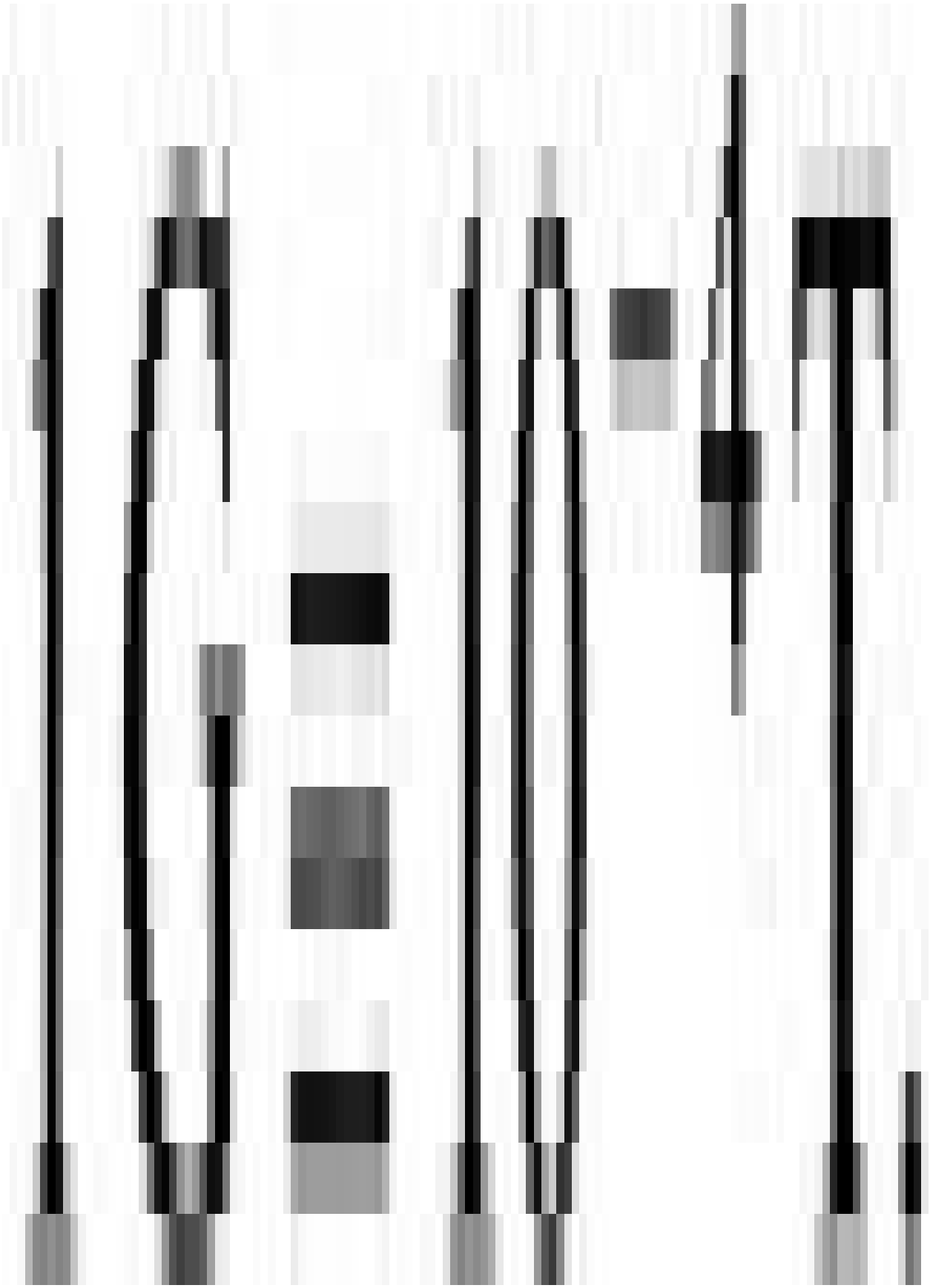
$$F_y = q(v_z B_x - v_x B_z)$$

$$F_z = q(v_x B_y - v_y B_x).$$

The definition (4.15) shows that the dimensions of the field B are [mass] [charge]−1 [time]−1. The unit of magnetic field is such that a charge of one coulomb moving with a speed of 1 ms−1 perpendicularly to unit field experiences a force of one newton. This unit is called the tesla (T). The strengths of magnetic fields are often given in an alternative unit, called the gauss (G), which is ten thousand times smaller than the tesla,

Figure 4.5. The direction of the force on a positively charged particle moving in a magnetic field.


v
F
B
F
v
B



A magnetic field of one gauss is comparable with the Earth's magnetic field. More precisely the field at the surface of the Earth varies from about 0.3 gauss at the equator to about 0.6 gauss at the poles. Typical iron-cored electromagnets can produce steady fields of up to about 2 T, while air-cored electromagnets can generate fields of 100 T for a short time over volumes of a few cubic cm. On a microscopic scale fields as large as 10^4 T have been found to occur near the nuclei of certain highly ionized atoms.

If an electric field E is present as well as a magnetic field B, a charge q experiences an additional force qE and the total force is

(4.16)

This equation defines the fields E and B in free space, where the force on a moving test charge can be measured. The test charge must be small enough not to affect the charge and current distributions that give rise to the fields, otherwise the measured force will not give those fields present without the test charge.

The way we have attributed steady magnetic fields to steady currents is analogous to the way we ascribed steady electric fields to stationary charges. The electrostatic field was defined in Equation (1.4) via the inverse square law. A similar approach to the discussion of steady magnetic fields due to currents would be to define the field B of a set of moving charges. We choose not to do this, but instead to define B by the Lorentz force law. This is because the magnetic field due to a single moving charge is rather complicated, and it is very difficult to deduce the magnetic fields generated by complex patterns of currents in conductors from this starting point. Later in this chapter we shall derive laws enabling us to calculate magnetic fields due to steady currents. These laws will eventually enable us to derive an expression for the magnetic field of a moving charge.

The Lorentz force law is found experimentally to be true even for particles moving at speeds close to the speed of light. The charge on a body is thus independent of its speed. All observers in uniform relative motion assign the same value to the charge on a body. Electric charge is therefore relativistically invariant just as the rest mass of a body is invariant in relativistic mechanics. We shall find later, in Chapter 14, that the laws of electromagnetism which we will derive require no modification to satisfy the principle of special relativity.

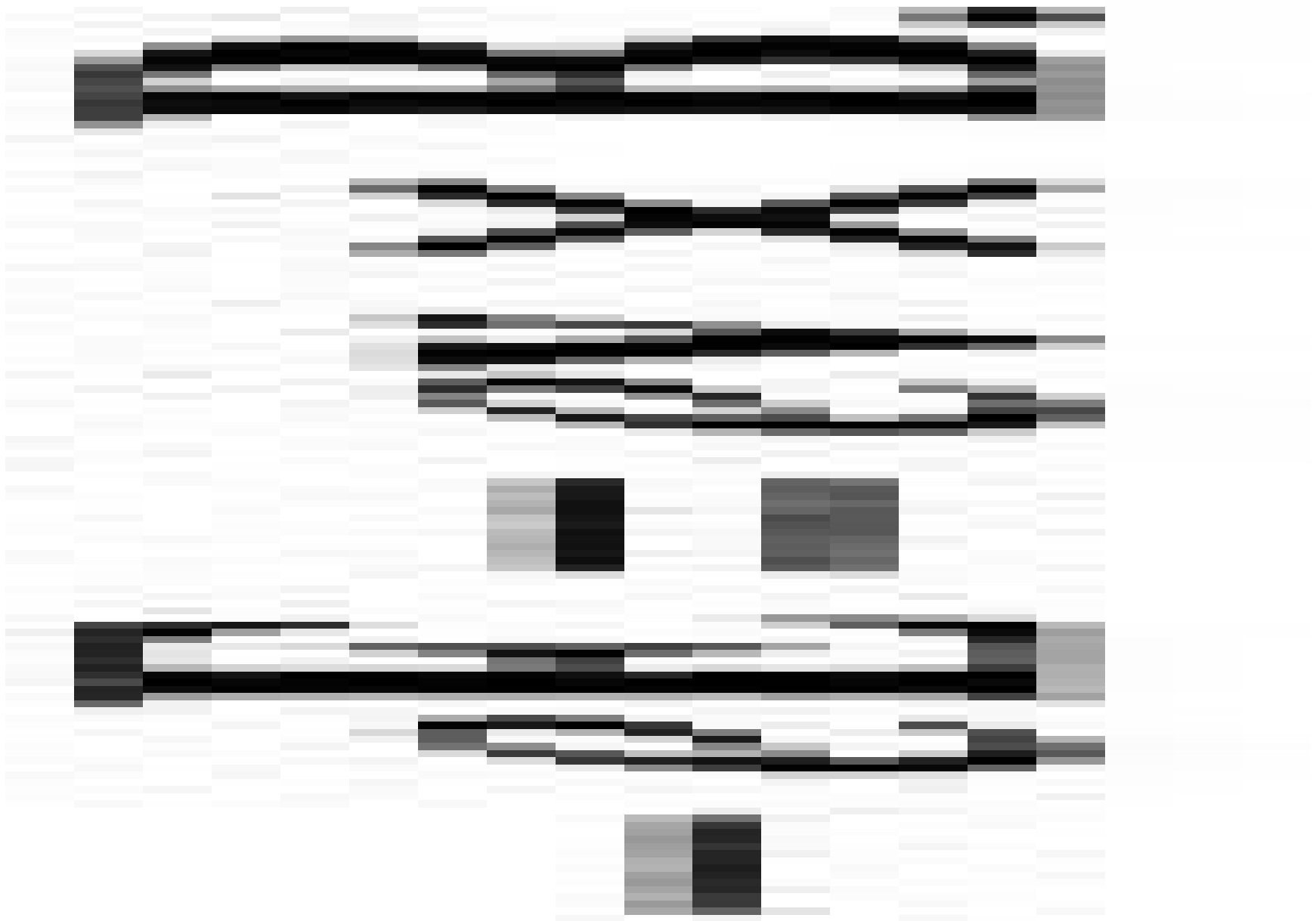
4.2.2 Magnetic field lines

The magnitude and direction of a magnetic field can be represented on a diagram by field lines, in the same way as an electrostatic field can be shown by lines of force. The direction of the magnetic field at any point is given by the direction of the field line at the point and in a sense given by an arrow on the line; the magnitude of the field is proportional to the density of field lines near the point.

One approach to the determination of the properties of the magnetic field, and hence its relationship to the currents that produce it, is to determine experimentally the fields produced by simple current distributions. The measurement of the Lorentz force on a moving charge provides (at least in principle) a method for determining magnetic fields. The direct measurement is usually not practical, but an indirect measurement, based on an effect known as the Hall effect, is often used. We will outline the theory of the Hall effect meter before discussing the magnetic fields that arise in a few simple examples.

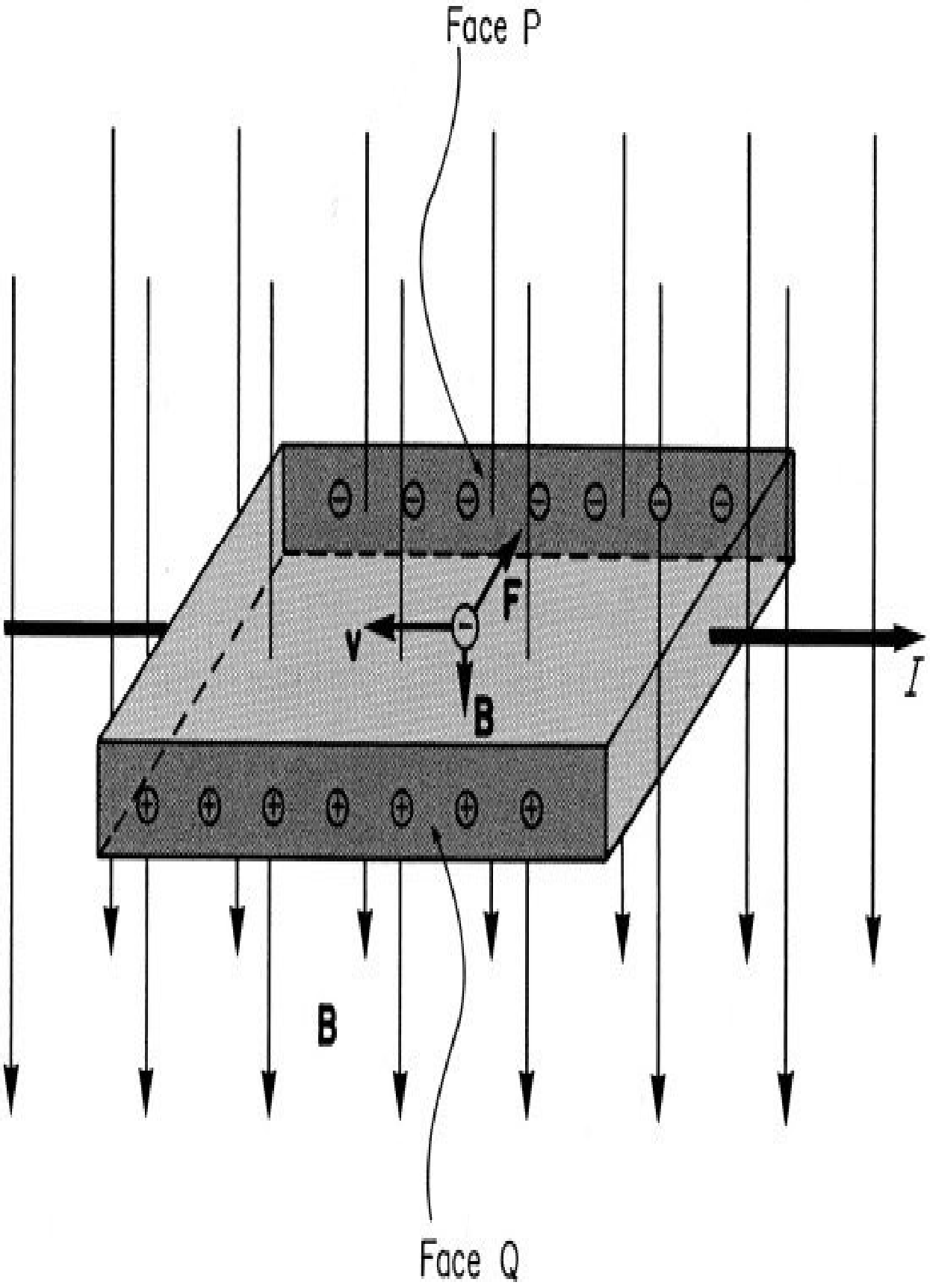
In a Hall effect meter there is a current I flowing in a thin conducting rectangular bar, as shown in Figure 4.6. Suppose that the current, from left to right, corresponds to the net motion of electrons from right to left with mean drift velocity v. If there is a uniform field B acting over the bar in the direction

shown, the electrons will experience a force F equal to –ev×B which tends to move them over towards the face P of the bar. Electrons will be forced towards this face until a charge distribution is set up which generates an electric field E just big enough to prevent more electrons from accumulating on the side near face P. The electric force – eE on the electrons is then equal and opposite to that due to the magnetic field,

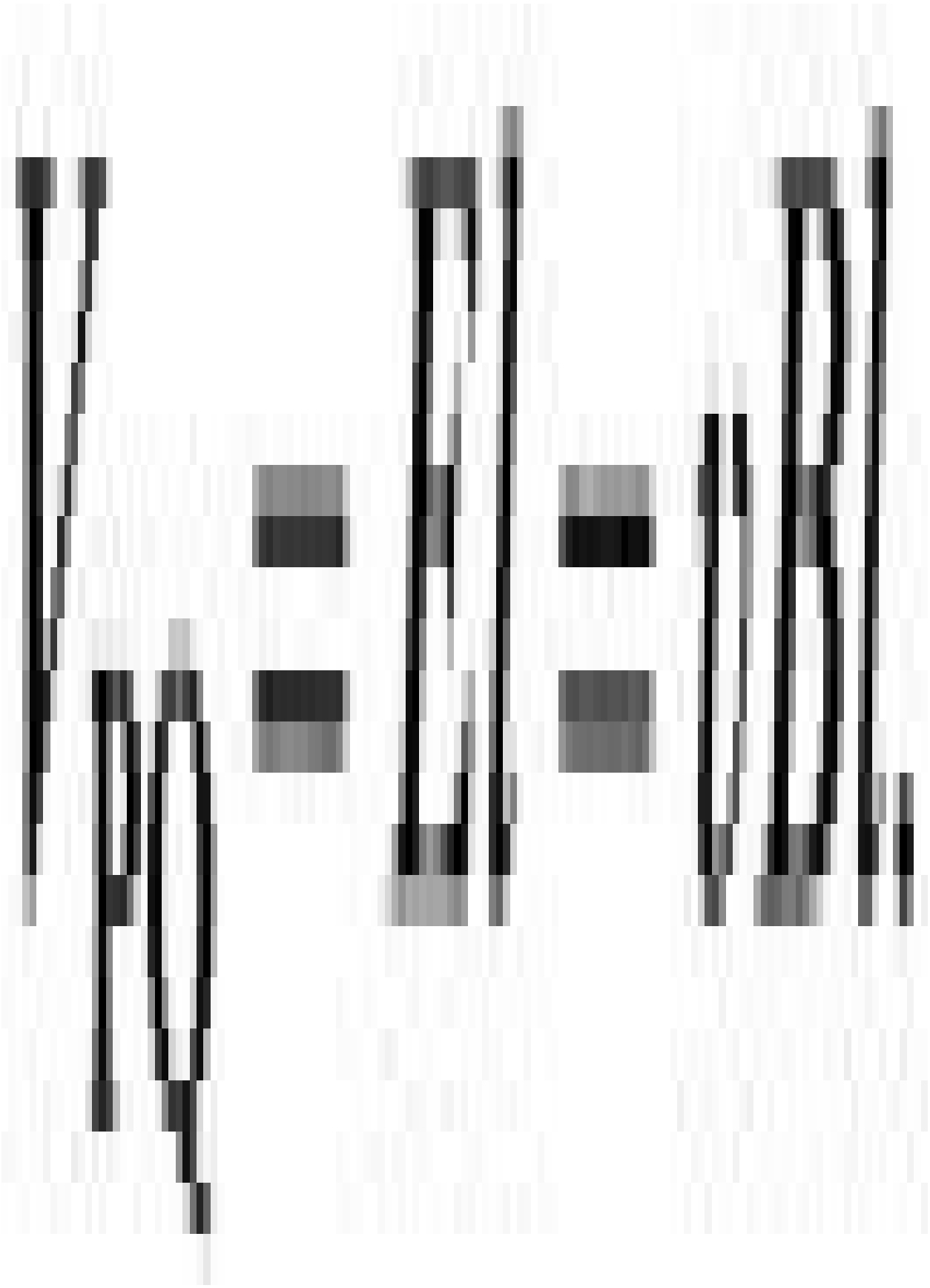


The occurrence of the transverse electric field is known as the Hall effect. The electric field in the bar has constant magnitude υB, and is in the direction from face Q to face P. If l is the distance between the faces P and Q, the potential

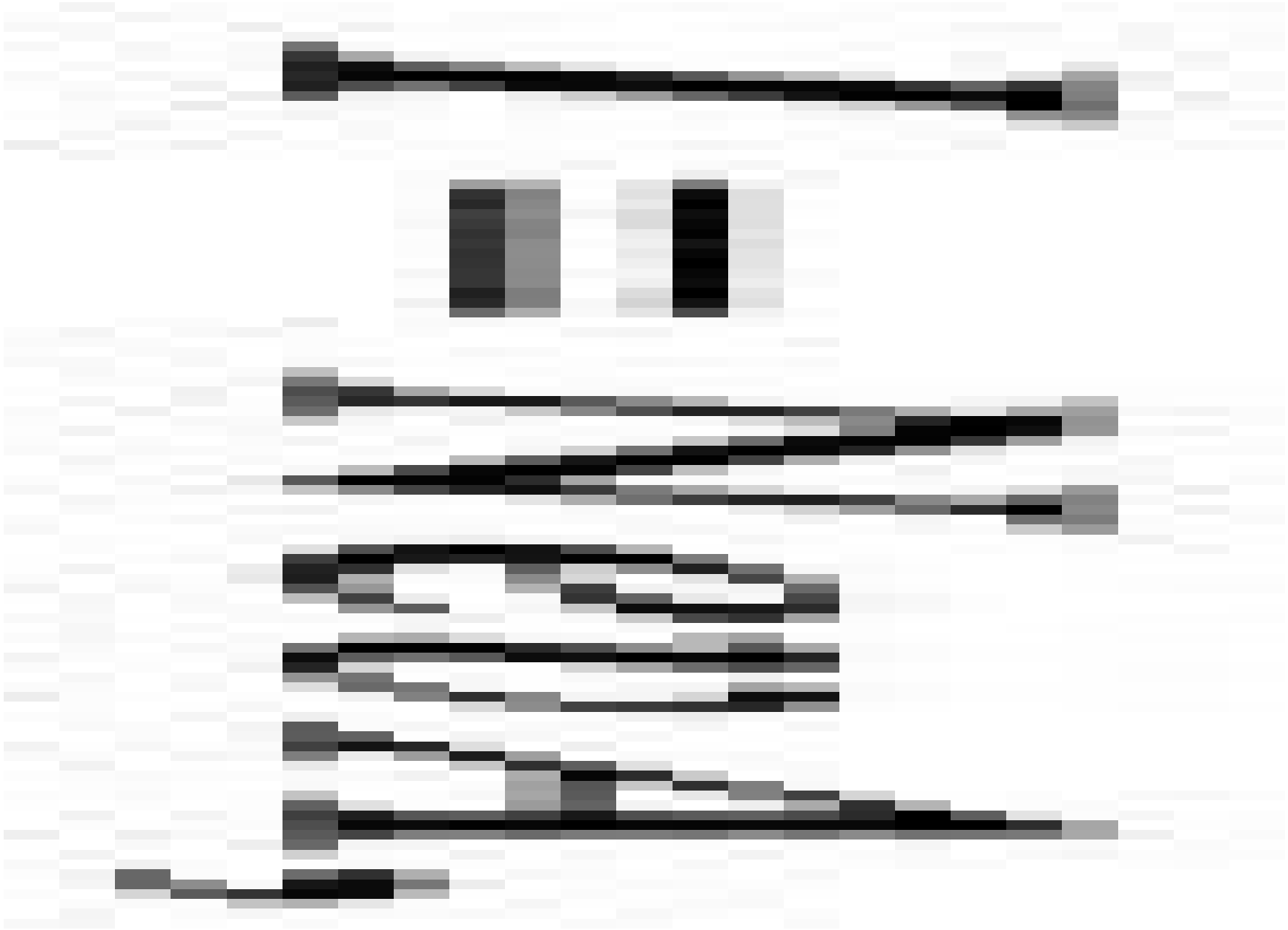
Figure 4.6. If a current flows in a conducting bar placed in a magnetic field, charges are produced on the faces P and Q as shown. This charge distribution gives rise to a uniform transverse electric field.


Face P
F
v
B
I
B
Face Q

VpQ between the faces is



The current I is given by Equation (4.3)



where N is the number of conduction electrons per unit volume, and A is the cross-sectional area of the bar. The potential difference between the faces is thus

(4.17)

A measurement of the potential difference across the bar thus gives the magnitude of the field B if the other quantities in Equation (4.17) are known.

Let us examine a common metal like copper in order to see whether it would be a suitable candidate for use in a Hall effect meter. If the bar were 0.5 cm wide (l= 0.5 cm), and 0.1 cm thick (A = $5\times 10-6m^2$), and carried a current of 1 A, Equation (4.17) predicts a value of the potential difference VPQ equal to about $10-7$ V for a field of one tesla, assuming that there is one conduction electron for every copper atom. The effect in copper is thus too small to be generally useful. Hall effect meters usually incorporate semiconducting materialsin which there are of the order of 10^7 times less conduction electrons per unit volume, and hence according to Equation (4.17) a proportionately larger potential difference. A typical commercial instrument contains a piece of semiconductor a few mm wide, having a carrier density of the order of 10^{15} electrons per cm^3, and passing a current of a few mA.

Using a Hall effect meter consisting of a thin slice of semiconductor we may now investigate the magnetic field lines around some simple current distributions. Figure 4.7 shows the field due to a steady current in a long straight wire. The field is cylindrically symmetrical, and its magnitude is inversely proportional to the perpendicular distance from the wire. The direction of the field at any point is perpendicular to both the wire and the perpendicular from the point to the wire. The field lines are circles, and the direction of the arrows on the circles is given by the right-hand screw rule. This rule says that the circles go round in the direction the threads of a right-hand screw would rotate if it were

screwed in the direction of the current.

Figure 4.8 shows the magnetic field lines from a small loop of wire carrying a current. (The leads carrying the current into and out of the loop are close together so that their magnetic effects cancel. They are not drawn.) The field from the small current loop is axially symmetrical, i.e. it is independent of the azimuthal angle with respect to the axis of the coil. This means that a map of the field lines would look the same for any plane containing the axis of the coil. An important result from such field mapping exercises is that at distances from the loop large compared with the dimensions of the loop, the magnetic field varies in the same way as the electric field of an electric dipole.

There is one obvious feature of the field lines on Figures 4.7 and 4.8 that is shared by the magnetic field lines arising from any current distribution. The lines of the field B are continuous. Unlike electrostatic field lines, lines of the field B have no beginning and no end. This means that there are no free ‘magnetic charges’ or ‘magnetic poles’, a fact which has important consequences for all of electromagnetism. Let us write this simple experimental fact in the form of a convenient mathematical equation. Consider a closed surface S around a volume V. Since the lines of the field B are continuous, as many lines enter the volume V as leave it. This means that the total outward flux of the field B over the surface S is zero,

Figure 4.7. The magnetic field of a current in a long straight wire.

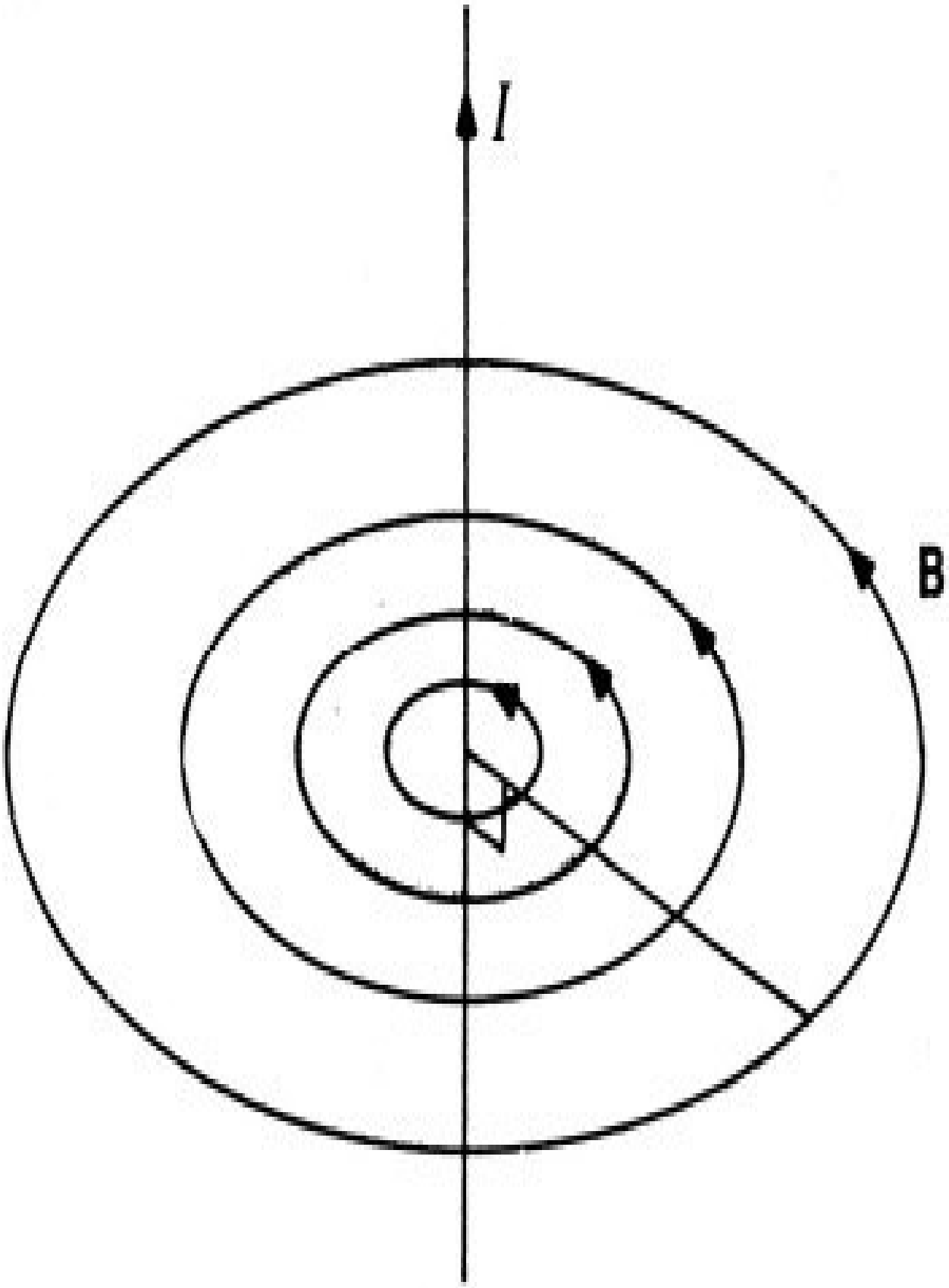
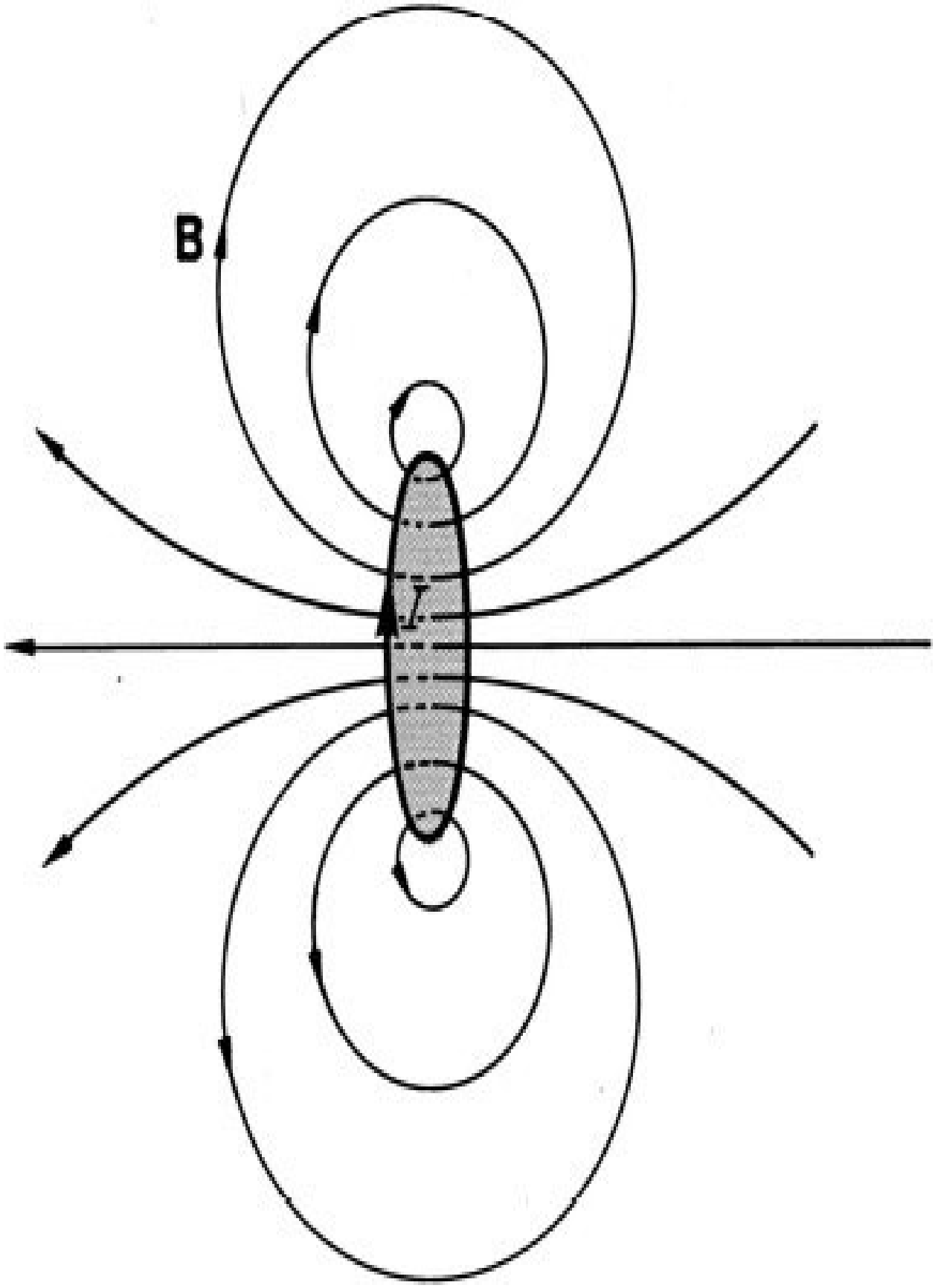
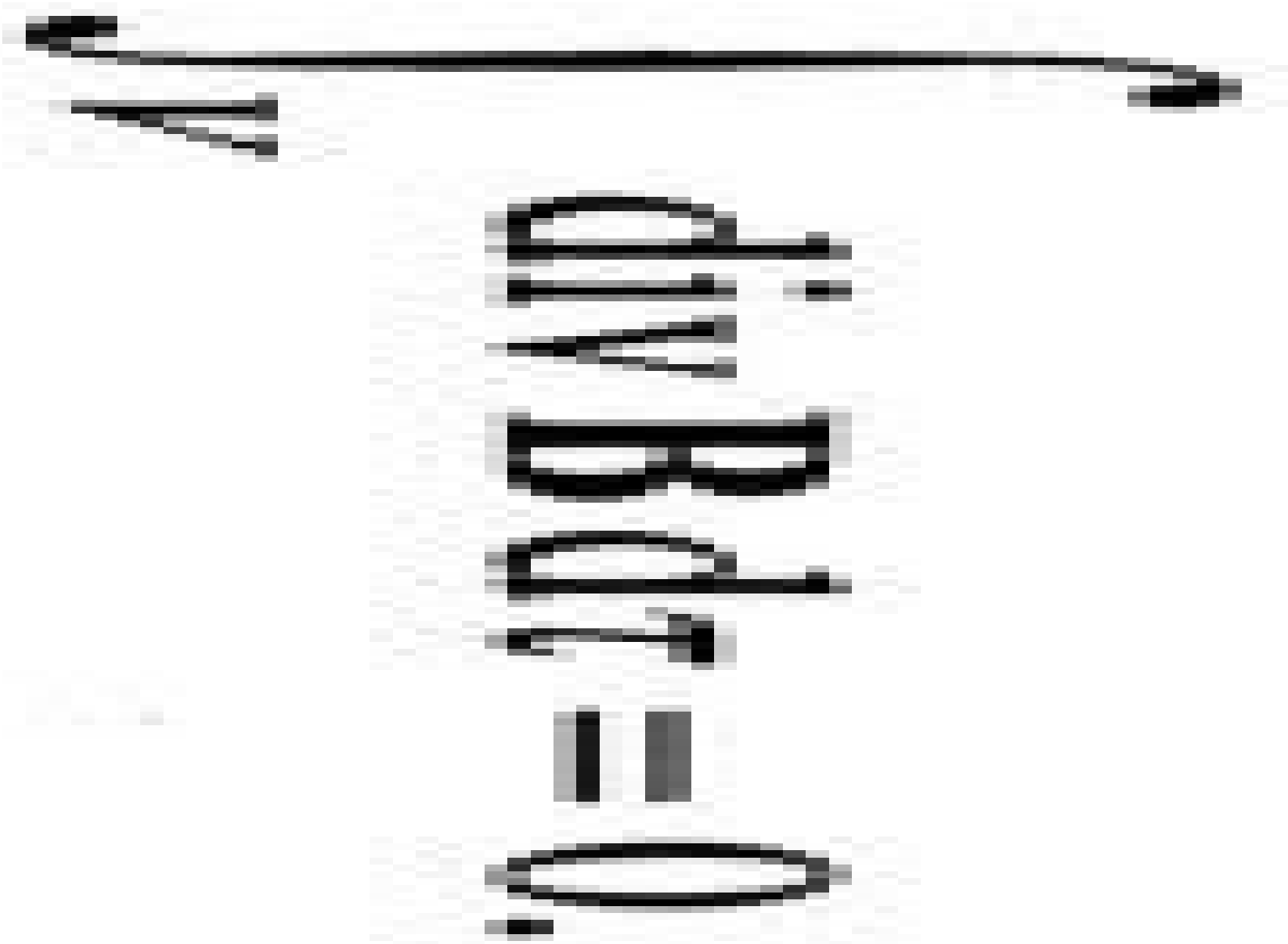

I
B

Figure 4.8. The magnetic field near to a small circular loop of wire carrying a current.


B
I

(4.18)

We may write this equation using the divergence theorem, Equation (1.21), to give



This equation is true whatever the size of the volume V, hence at all points the integrand must be zero, and everywhere

(4.19)

Equation (4.19) is one of the fundamental laws of electromagnetism, on the same footing as Coulomb's law. There have been no observed violations of this law that the divergence of the magnetic field B is everywhere zero; no free magnetic 'charges' have yet been discovered. The development of the theory of electromagnetism depends on this last negative observation. Since electromagnetic theory always gives the right answer we have very strong indirect evidence for the general validity of Equation (4.19).

4.3 THE MAGNETIC DIPOLE

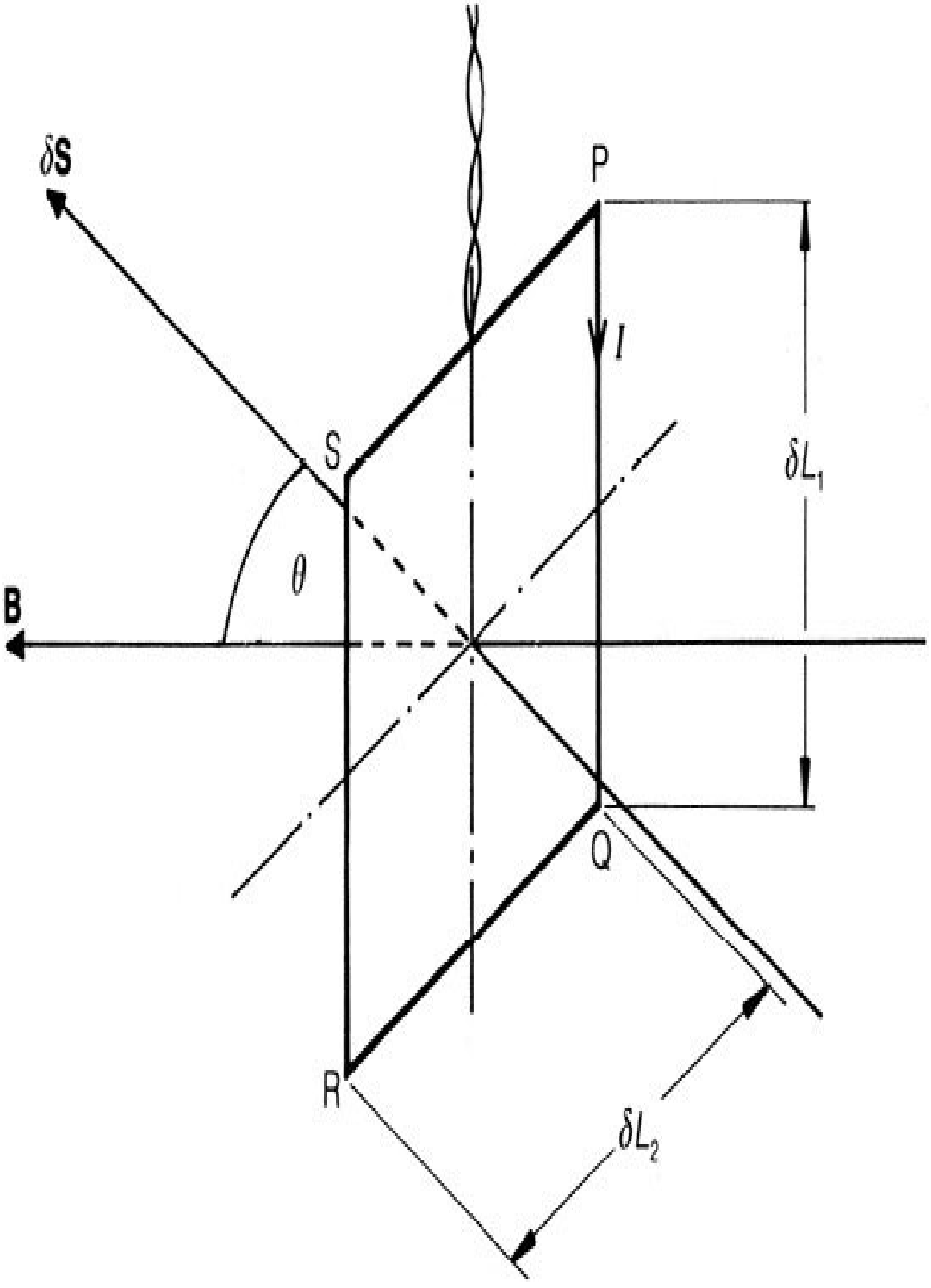
The whole of this section is devoted to a more detailed discussion of the magnetic properties of small current loops like the one in Figure 4.8. There are two reasons why the small current loop deserves such attention. First it will be seen later that we can evaluate the field generated by any circuit carrying a current in terms of the field due to a small loop. Secondly the circulating electrons in an atom are themselves small current loops and we shall make use of the results of this section when studying the magnetic behaviour of atoms.

4.3.1 Current loops in external fields

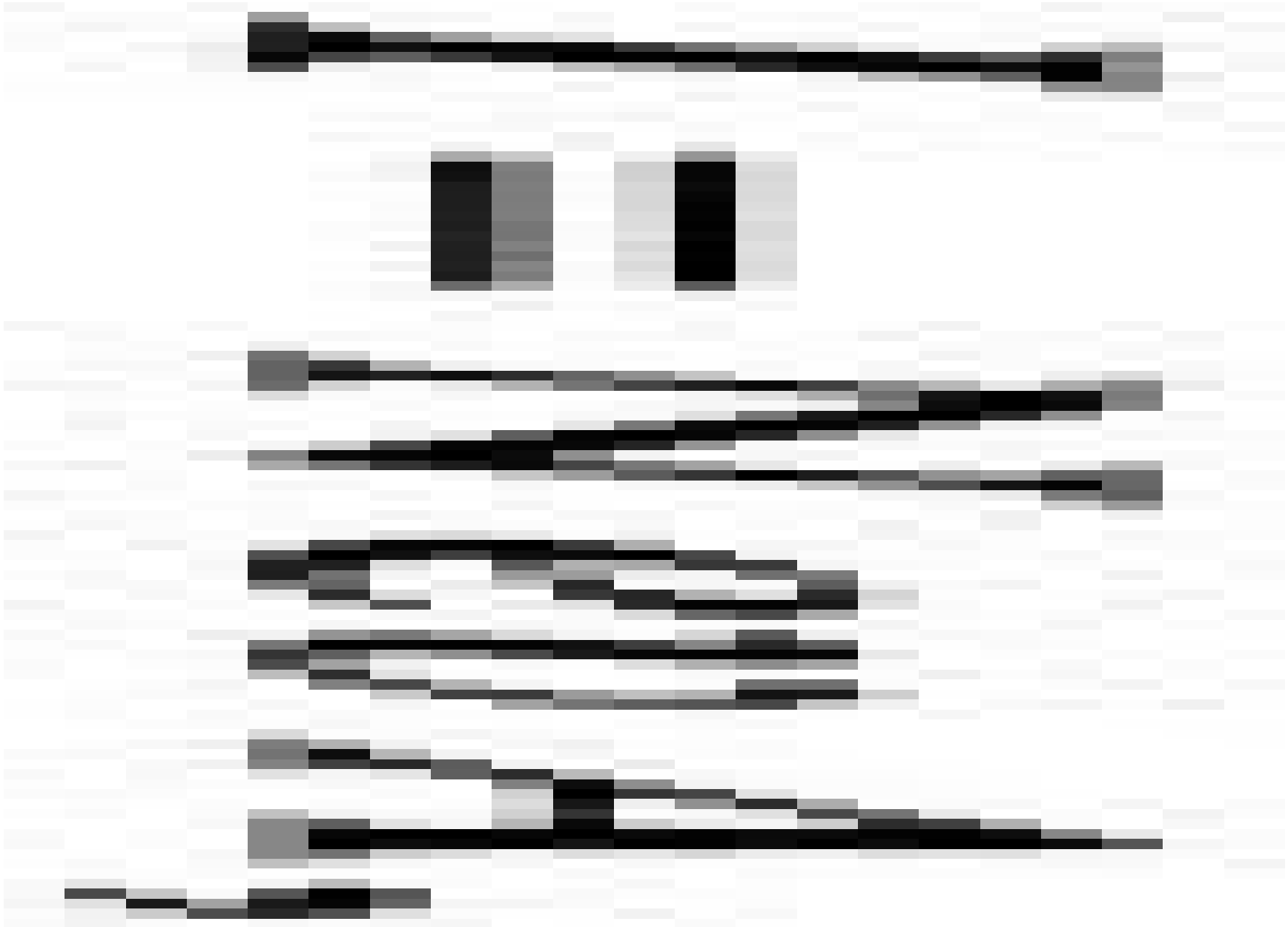
We begin by showing that a small loop carrying a current behaves in an external magnetic field rather like an electric dipole in an external electric field.

Consider the small rectangular coil shown in Figure 4.9, which is suspended in a vertical plane, and consists of a single loop of wire. (The input and output leads are wound close together so that we may neglect their magnetic effects.)

Figure 4.9. A small rectangular coil carrying a current, situated in a uniform magnetic field. The coil rotates about an axis perpendicular to the field direction.


δS
P
I
S
δL1
θ
B
Q
R
δL2

The coil is placed in a magnetic field B which is constant over the area of the coil. The field lines of B are horizontal, at an angle θ to the normal to the plane of the coil, as shown in the diagram. If the coil carries a current I and the wire has a cross-sectional area A, then



where N, e and υ have the same meanings as in Equation (4.3). In a small length dl of the coil on the side PQ there are AN dl electrons moving upwards with speed υ. The force on these is

(4.20)

where the vector dl is in the direction of the current along PQ. The total force δF1 on the side PQ thus has magnitude IδL1B, since the side PQ and the field B are everywhere perpendicular. This force is in the direction shown in Figure 4.10, which is a view looking down on the coil. There is a force δF2 of equal magnitude on the side RS of the coil, but it is in the opposite direction to the force δF2 Similarly there is a force acting downward on the side QR but this is balanced by an equal and opposite force acting upwards on the top side PS. The resultant force on the whole coil is thus zero. However, there is a net torque δT on the coil tending to twist it about the vertical axis. This can be seen from Figure 4.10 to have magnitude

(4.21)

where δS is the area of the small coil. The torque δT tends to twist the coil anticlockwise (looking down from above the coil) and so is in the upwards direction,

Figure 4.10. A view looking down on the rectangular coil, showing the directions of the forces δF1 and δF2 and the direction of the vector δS.

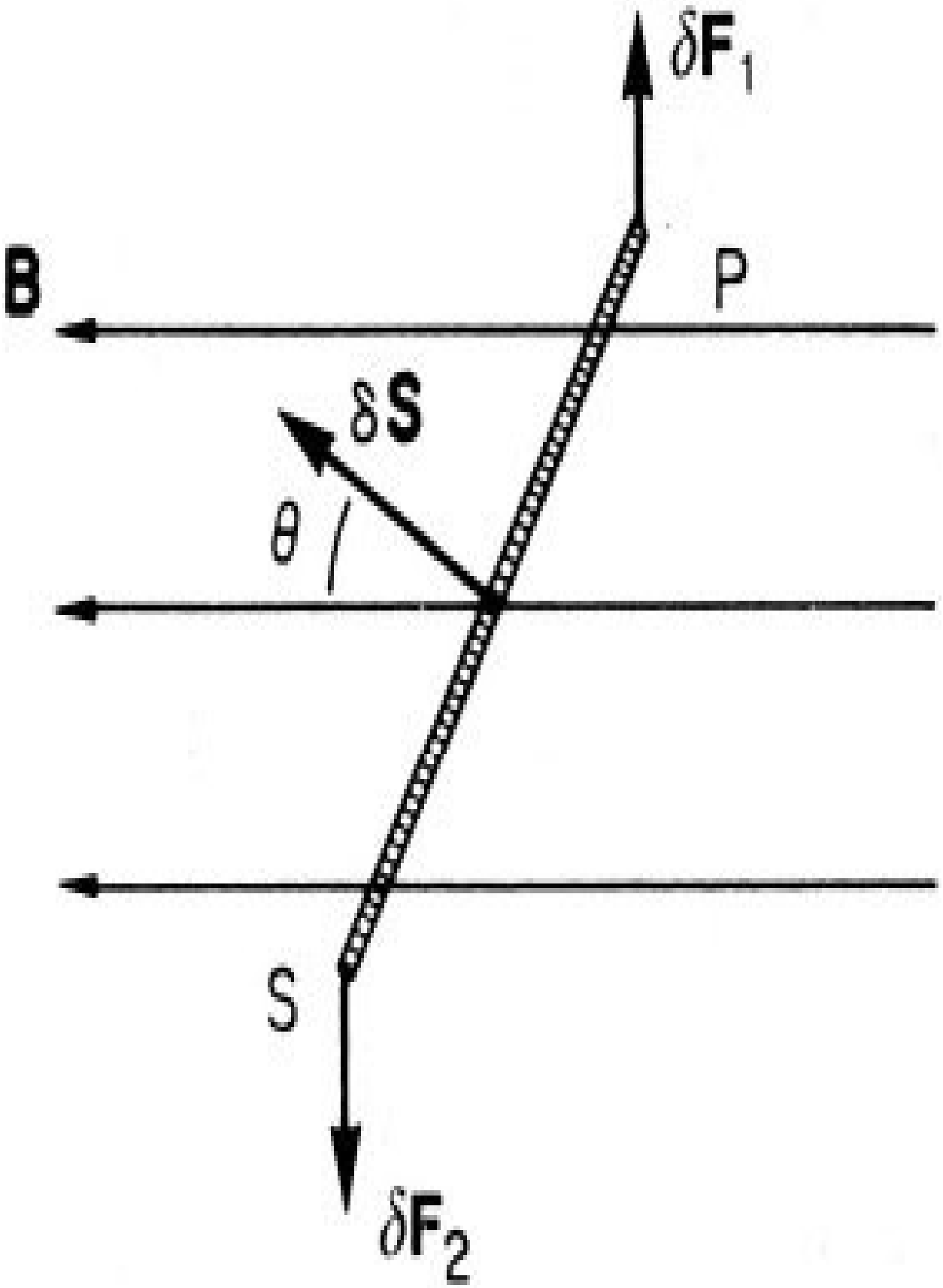
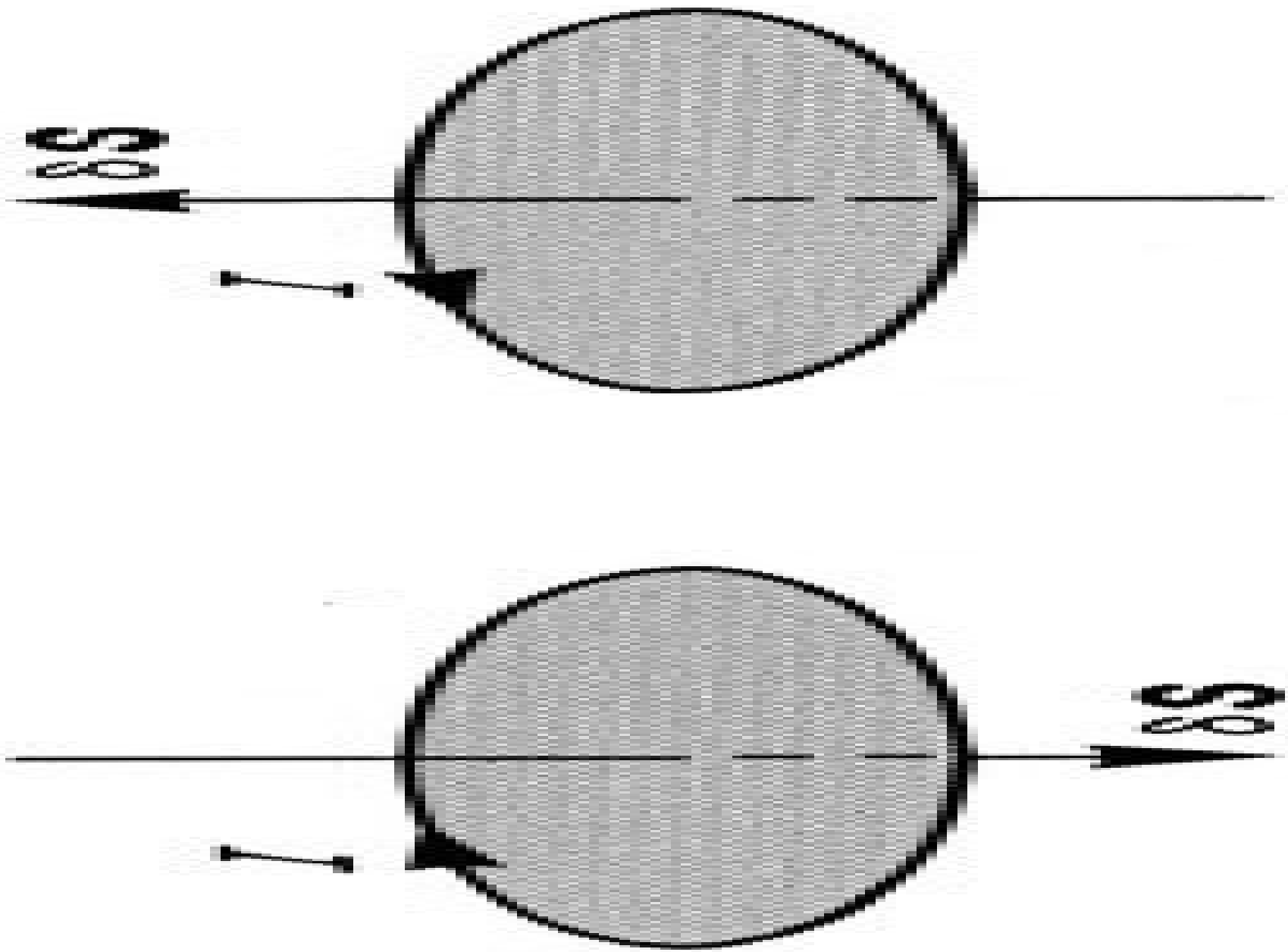

$\delta \mathbf{F}_1$
B
P
$\delta \mathbf{S}$
θ
S
$\delta \mathbf{F}_2$

Figure 4.11. The right-hand screw rule. The direction of the vector δS associated with the area of the loop is given by the direction in which a right-hand screw would advance if rotated in the same direction as the current in the loop.

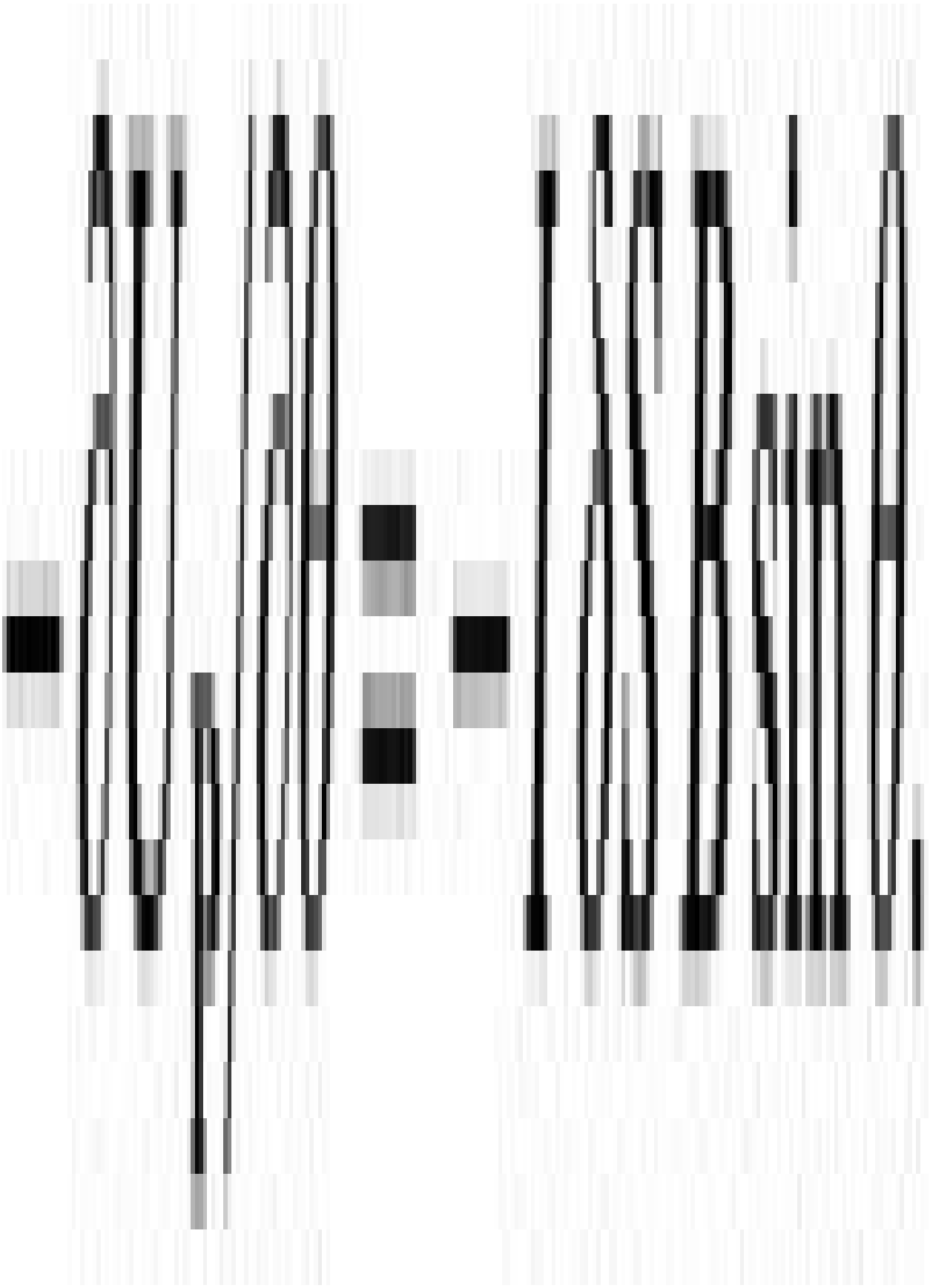


(4.22)

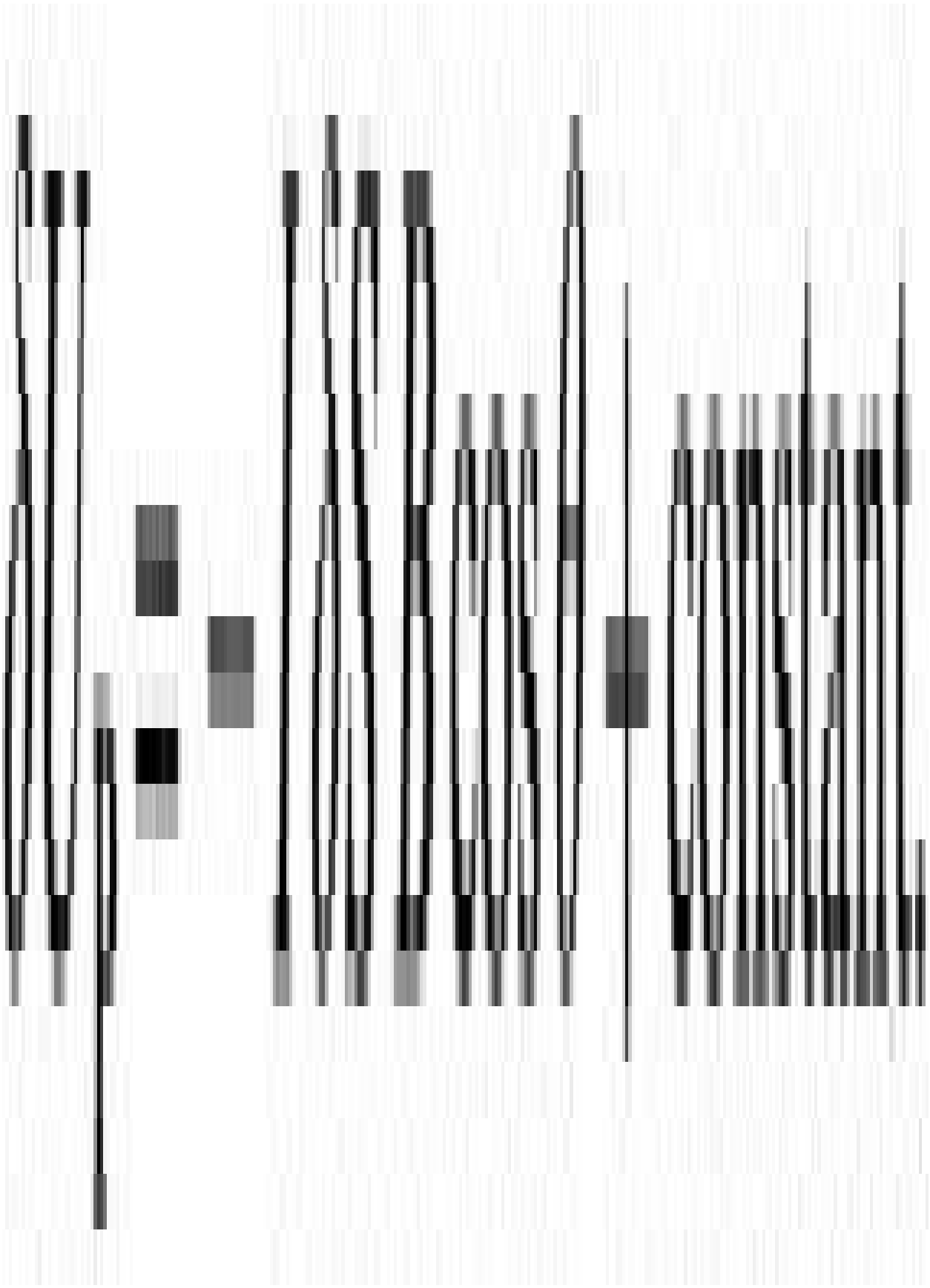
where δS is a vector along the normal to the plane of the coil. Its direction is given by the right-hand screw rule used with the direction of the current around the loop. For example Figure 4.11 shows a loop of current looked at from above; if the current is clockwise the vector δS is down; if the current is anticlockwise the vector δS is up. We have written the torque on the coil in the form of Equation (4.22) because it can be shown that this equation is true for a plane coil of any shape when placed in a uniform magnetic field.

In Equation (4.22) the field B is the external field present at the coil before the current was passed through it. The total field near the coil when it has the current I flowing in it is equal to the sum of the external field and the field arising from the current I. Now the force on a moving charge is related to the total field from all sources (other than the charge itself) at the position of the moving charge, so that it might seem that the derivation above needs to be modified. However, that part of the total field that arises from the current in the coil itself will cause no net force or torque on the coil. In the absence of an external field a coil through which a current is passed does not move or rotate. Hence the field B in Equation (4.22) can equally well be taken to be the total magnetic field.

Equation (4.21) enables us to calculate the potential energy UP of a current loop in an external field. The torque on the coil in Figures 4.9 and 4.10 tends to decrease the angle θ. In terms of the potential energy the torque tending to increase θ is $-\partial U_p/\partial\theta$. Hence from Equation (4.21)



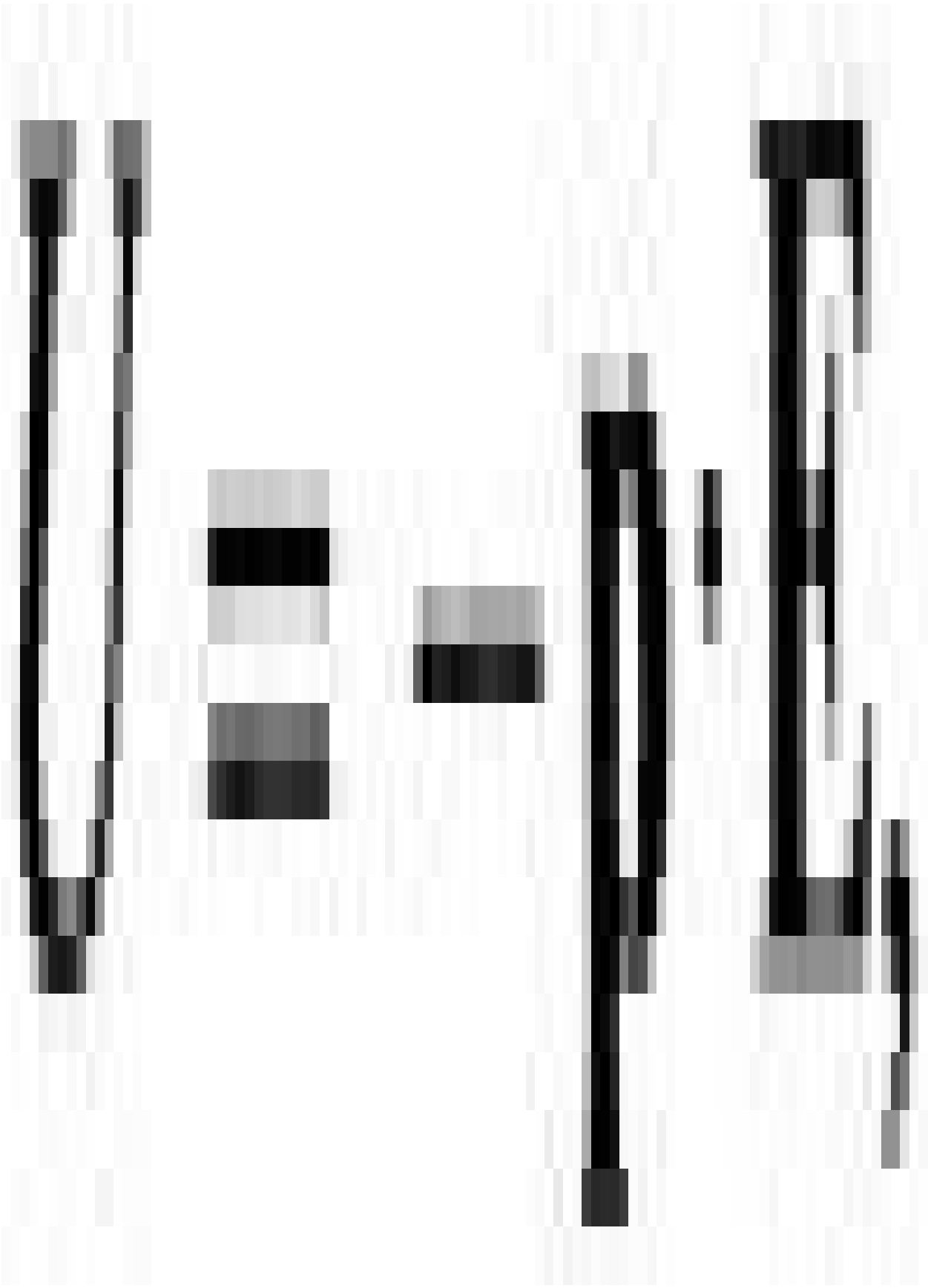
Integration of this equation gives



If we choose the constant so that the potential energy δU_p is zero when θ is equal to $\pi/2$ we have

(4.23)

By considering the torque exerted on the coil by a vertical component of the magnetic field, it can be shown that this equation holds whatever the direction of B. Equation (4.23) is similar in form to Equation (2.11),



which gives the potential energy of an electric dipole, of dipole moment p, in an electric field E.

4.3.2 Magnetic dipoles and magnetic fields

We have already remarked that the magnetic field produced by a small current loop has a spatial variation similar to that of the electrostatic field of an electric dipole away from the immediate vicinity of the loop. In the above section we showed that in an external field a small current loop behaves in the same way as an electric dipole in an electric field. Experiments also show that two small coils carrying currents exert forces and torques on each other which vary with distance apart and relative orientation of the coils in the same way as the forces and torques between two electric dipoles. The electromagnetic effects of a small current loop can thus be determined by associating a magnetic dipole moment with the loop.

We define the magnetic dipole moment m of a small coil of area S carrying a current I to be

(4.24)

where S is the vector perpendicular to the plane of the coil in a direction related to the current by the right-hand screw rule. In terms of the dipole moment, Equations (4.22) and (4.23) for the torque on a small coil and its potential energy in a field B become

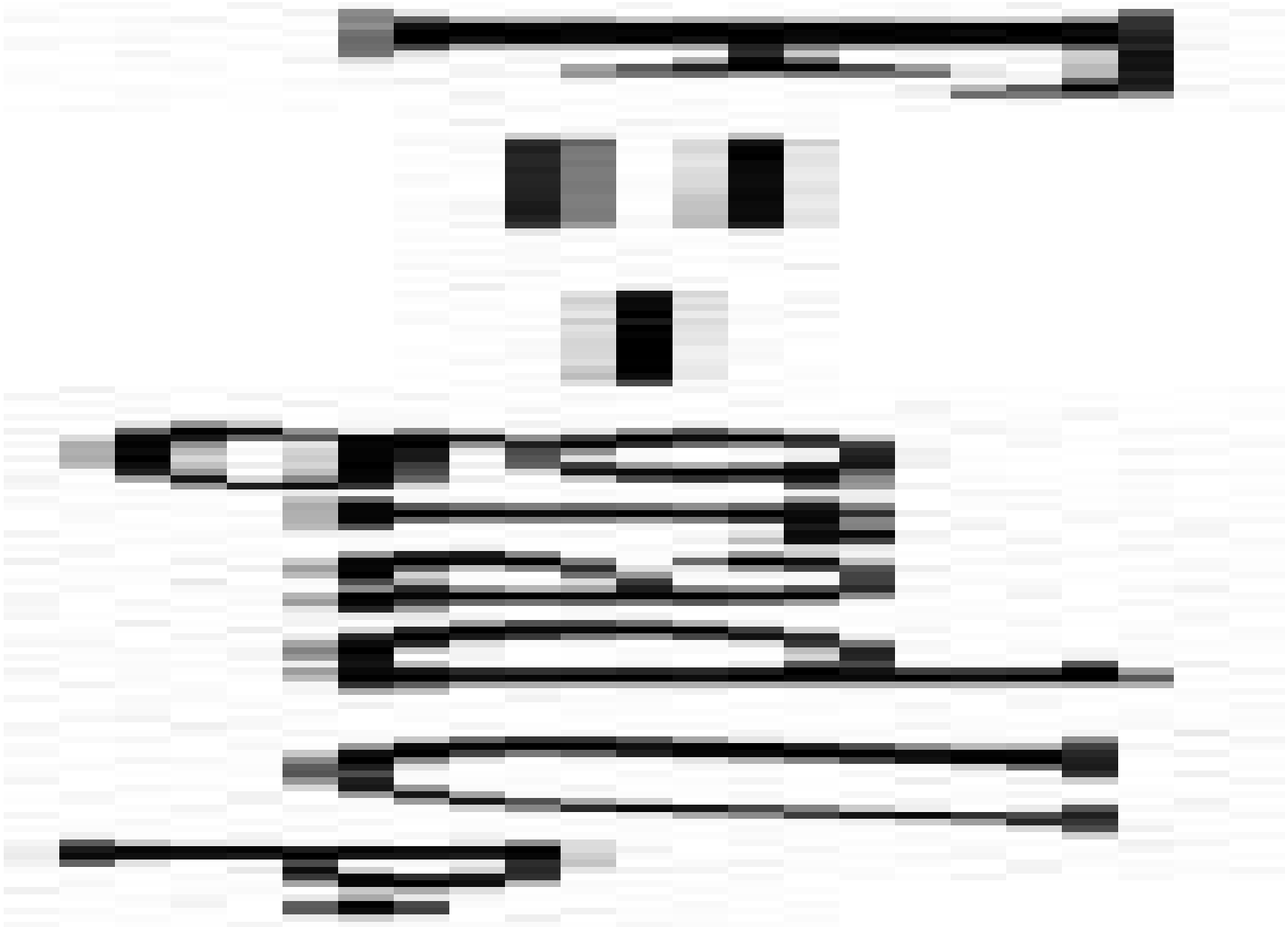
(4.25)

and

(4.26)

Magnetic dipole moments are measured in A m². A single loop of wire of area 50 mm² and carrying a current of 0.2 A, has a dipole moment of $m = 5 \times 10^{-5} \times 0.2 = 10^{-5}$ A m². If the loop were placed in a uniform magnetic field of 0.1 T with its magnetic moment perpendicular to the field direction, Equation (4.25) predicts that the torque on it is 10^{-6}Nm. This is roughly equivalent to the torque produced by two forces each equal to the weight of 10^{-4}g at the surface of the earth acting over a lever arm of length 1 m.

Equation (4.26) may be used to calculate the force on an atomic dipole in a non-uniform magnetic field. This force is given by



Hence

(4.27)

and if the dipole points in the z-direction the z-component of the force on it becomes m∂Bz/∂z. In those atoms which possess a magnetic moment, the magnitude of the moment is typically about 10−23 A m², corresponding classically to a current of about 3×10−4A in a loop of atomic radius. The maximum field gradients attainable in the laboratory over useful distances are about 10 T m−1 An atom in such a non-uniform magnetic field will experience a force of about 10−22N. This is about fifty times the weight of an atom from the middle of the Periodic Table and can thus be measured without great difficulty.

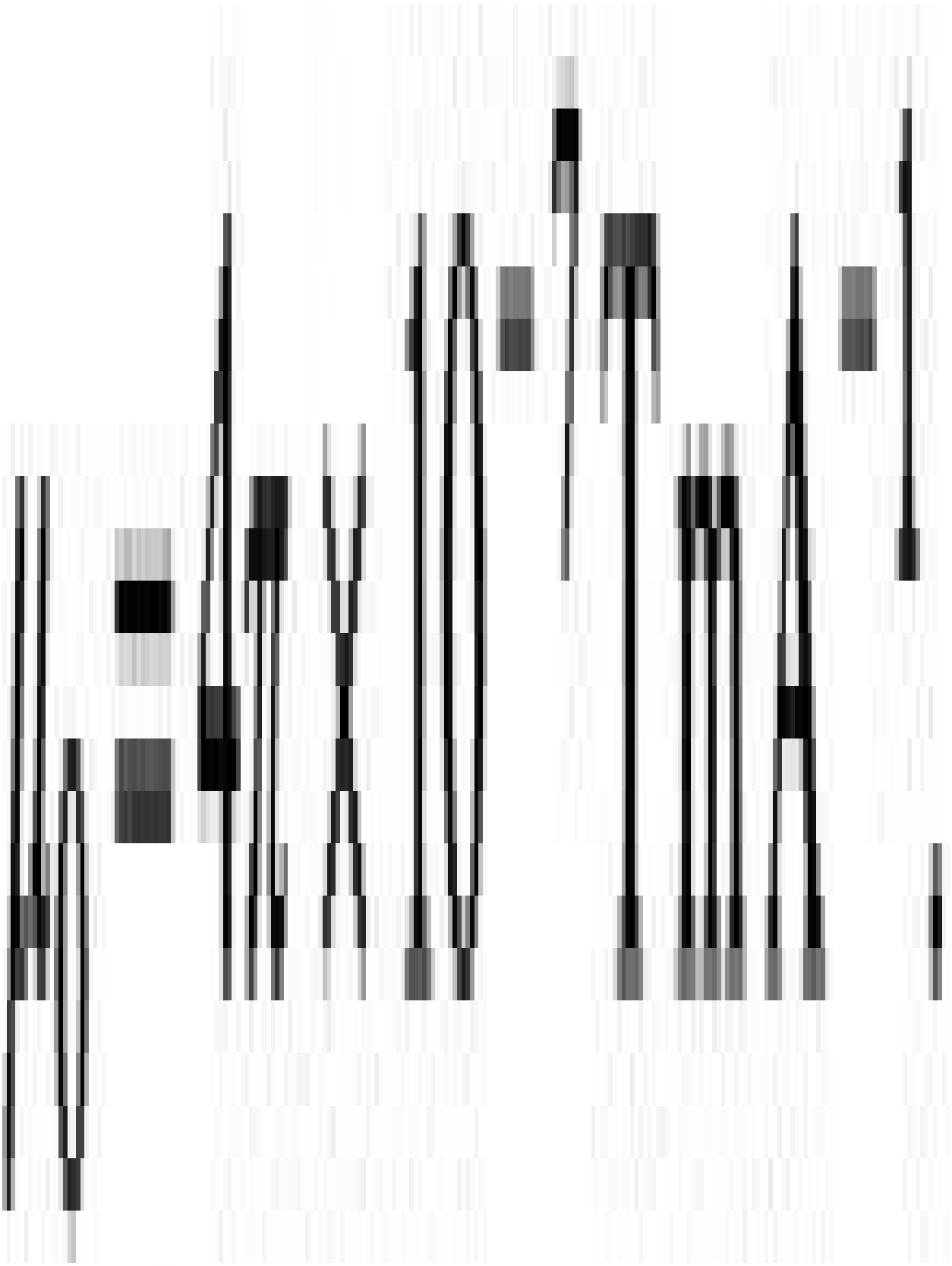
The dependence of the field B of a small current loop on position is the same as that of the field of an electric dipole, which can be calculated as outlined in section 1.5.3. For example, on the axis of the loop at large distances r, Equation (1.34) indicates that

(4.28)

Let us now examine the constant of proportionality required in this equation, which we write as the combined constant μ0/4π in order to simplify the form of some important equations. This gives

(4.29)

The constant μ_0 is called the permeability of free space. The right-hand side of Equation (4.28) has the dimensions [current] [length]$^{-1}$, whilst the dimensions of the field B are, from Equation (4.15), [mass] [current]$^{-1}$ [time]$^{-2}$. Thus the constant μ_0 has the dimensions [mass] [length] [charge]$^{-2}$. The numerical value of the constant cannot be chosen arbitrarily, since the units of all other quantities in Equation (4.29) have already been determined. The numerical value of μ_0 in SI units is



As one may guess, the convenient value for the constant μ_0 is no accident. In the SI system of units the permeability of free space is chosen to be exactly equal to $4\pi \times 10^{-7}$. The hierarchy of electrical units is discussed in detail in Appendix A.

Using the value $4\pi \times 10^{-7}$ in Equation (4.29) we can determine the field on axis produced by the current of 0.2 A in the loop discussed earlier in this section. At a point 1 m from the loop the field is equal to 2×10^{-12} T.

4.4 AMPÈRE'S LAW

In this section we proceed from very small current loops to more realistic circuits. We investigate the behaviour of the magnetic field of a current carrying coil in order to derive an expression relating the line integral of the field B around a closed loop to the current that passes through the loop. This integral relationship is called Ampère's law. It is analogous to Gauss' law, Equation (1.14), in electrostatics, and can be used to determine the magnetic fields due to some simple symmetric current distributions. Ampère's law is one of the most important laws of electromagnetism, and we shall see later that it can be expressed in a differential form which enables one to determine the fields of more complicated current distributions.

The magnetic field due to currents in circuits can also be obtained using the Biot–Savart law, which gives the fields as the sum of contributions from the small current elements which together comprise the circuit. The Biot–Savart law can be deduced from Ampère's law and vice versa.

Ampère's law is introduced in the next section, followed by a discussion of the Biot–Savart law and some examples of the calculation of magnetic fields.

4.4.1 The field of a large current loop

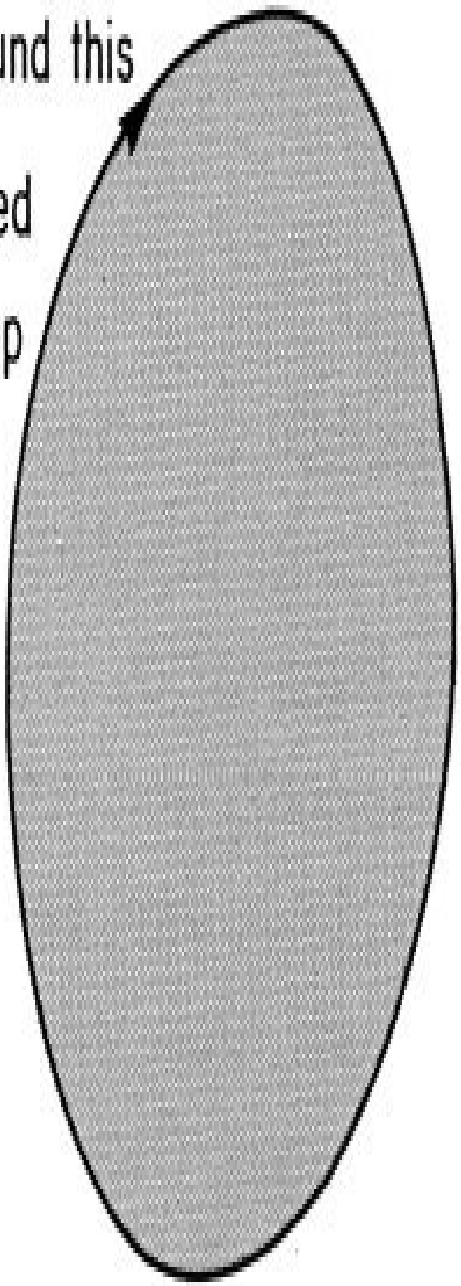
We will now examine the field of a current-carrying coil in more detail. We will first introduce a method for calculating the field near to the coil. This method considers the coil to be a sheet of magnetic dipoles, and a similar method can be used to determine the electrostatic field of a layer of electric dipoles. By discussing the physical differences between the two situations we derive an

expression for the line integral of the field B over a closed path threading the coil.

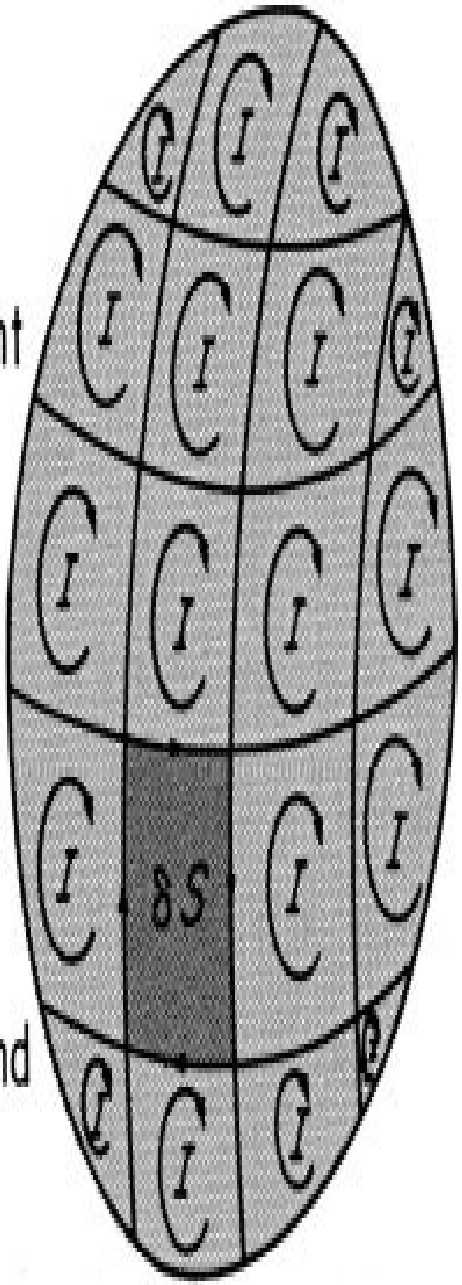
At points near to a coil which carries a current, we cannot treat the whole coil as one dipole, but we can consider it to be made up of very many dipoles by introducing the idea of the current grid. Figure 4.12 shows a loop of wire with a contour which we shall denote by the symbol s. Consider the surface S whose rim is the contour s and imagine this surface divided by a fine grid into very many small surfaces δS. If the same current I flows clockwise around the perimeters of all these small surfaces, the currents cancel except at the edge of the grid and the net result is a current I flowing clockwise around the contour s. Thus, for the purpose of calculating its magnetic effects, a current in a flat coil can be replaced by the same current flowing around the perimeters of the very many small surfaces that together make up the surface area of the grid. If we make the grid become more and more fine, the areas δS tend to become vanishingly small elemental areas dS. Each one of these can be considered to give rise to a dipole field at any point, no matter how near the point is to the coil.

Figure 4.12. The equivalence of a loop of wire carrying a current and a current grid. The current I around the perimeter of an area like δS inside the grid is cancelled by the currents around the perimeters of surrounding areas. The result is a current I around the original loop.

Current I around this closed loop



equivalent to currents I around many small loops covering area of large loop.



Let the direction of the current I around the perimeters of the areas dS define the direction of vectors dS via the right-hand screw rule. The current grid is equivalent to a sheet of elementary magnetic dipoles, each of moment dm = I dS. The magnetic field at any point can now be calculated from the sum of the fields of these elementary magnetic dipoles.

We may compare our problem, which we have reduced to that of determining the field of a uniform sheet of magnetic dipoles, to the one where we want to know the electrostatic field of a uniform sheet of electric dipoles. The fields outside an electric or a magnetic dipole have the same spatial variation. Hence the magnetic field on the right hand side of the current loop, for example, has the same dependence on position as the electric field on the right-hand side of a sheet of electric dipoles with similar size and shape to the flat surface S.

Let us consider the electrostatic field of a flat sheet of electric dipoles which has N dipoles per unit area each of moment p in a direction perpendicular to the sheet. An elemental area dS of the sheet contributes an amount $d\phi$ to the electrostatic potential at a point P, where

from Equation (1.31). Here r is the position vector of the point P with respect to an origin located at the position of the area dS, and the vector dS is in the direction of the dipoles p. The total electrostatic potential from the whole sheet is

(4.30)

The surface integral can be evaluated in a straightforward way on the axis of a circular sheet of dipoles. The axis is drawn from left to right on Figure 4.13, and we take the dipoles to point to the right. Beginning on the left-hand side of the sheet, the value of the integral changes smoothly from zero at the point a very far to the left to -2π at a point very near to the sheet. Beginning far on the right, the value of the integral changes smoothly from zero at the point b very far to the right to $+2\pi$ very close to the sheet. Over a path on axis, beginning at large distance on the left, the total electrostatic potential from the charge distribution goes from zero to $-Np/2\varepsilon 0$ just to the left of the sheet: it then increases to $+Np/2\varepsilon 0$ as the path continues across the gap, however small, between the layers of positive and negative charge which constitute the dipole sheet: it then decreases back to zero at a point a large distance to the right. The potential $\phi(r)$ is related to the field E by Equation (1.24), and taking the potential to be zero at the point a very far to the left of the coil where we will begin the line integral we have

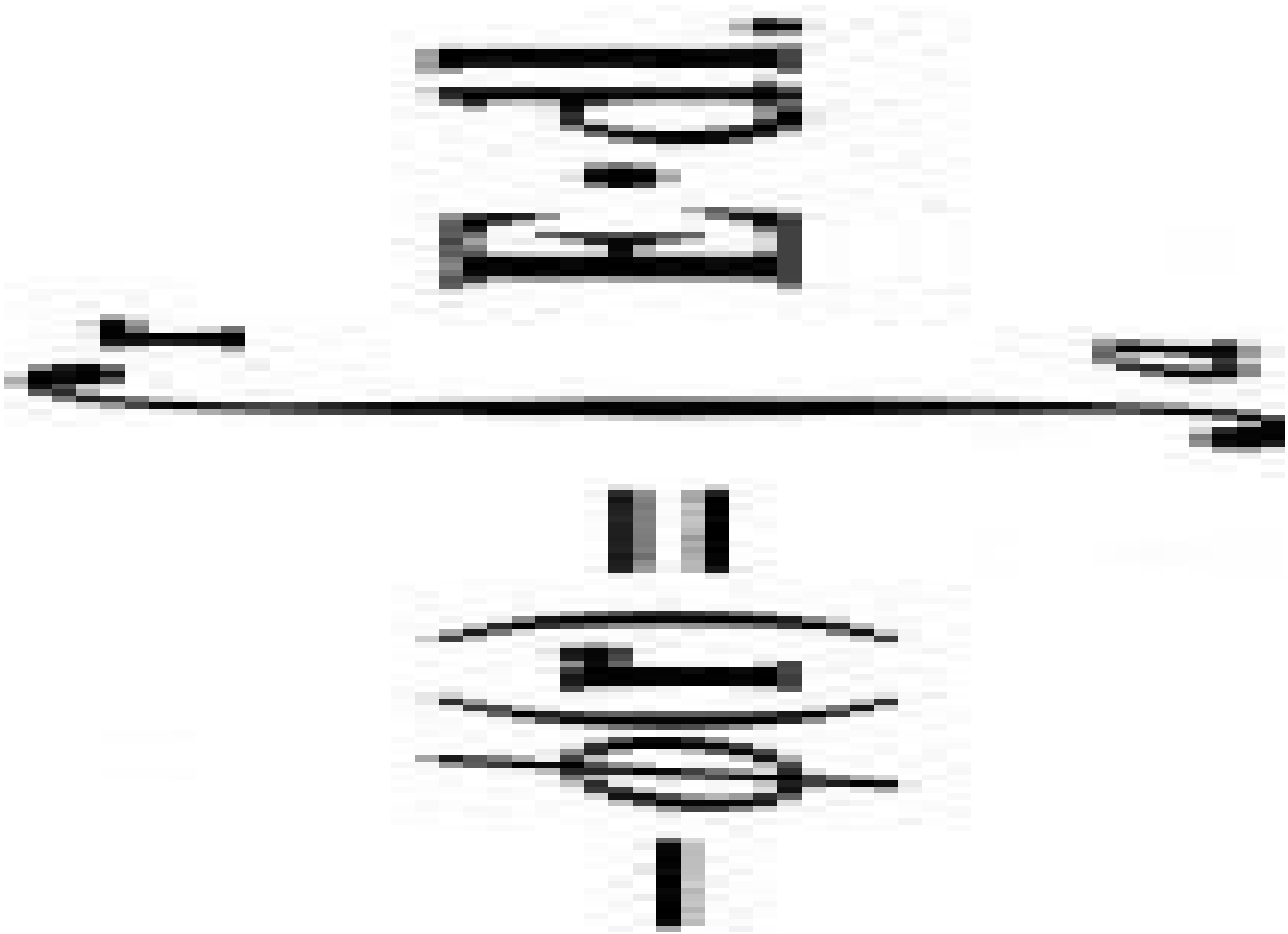
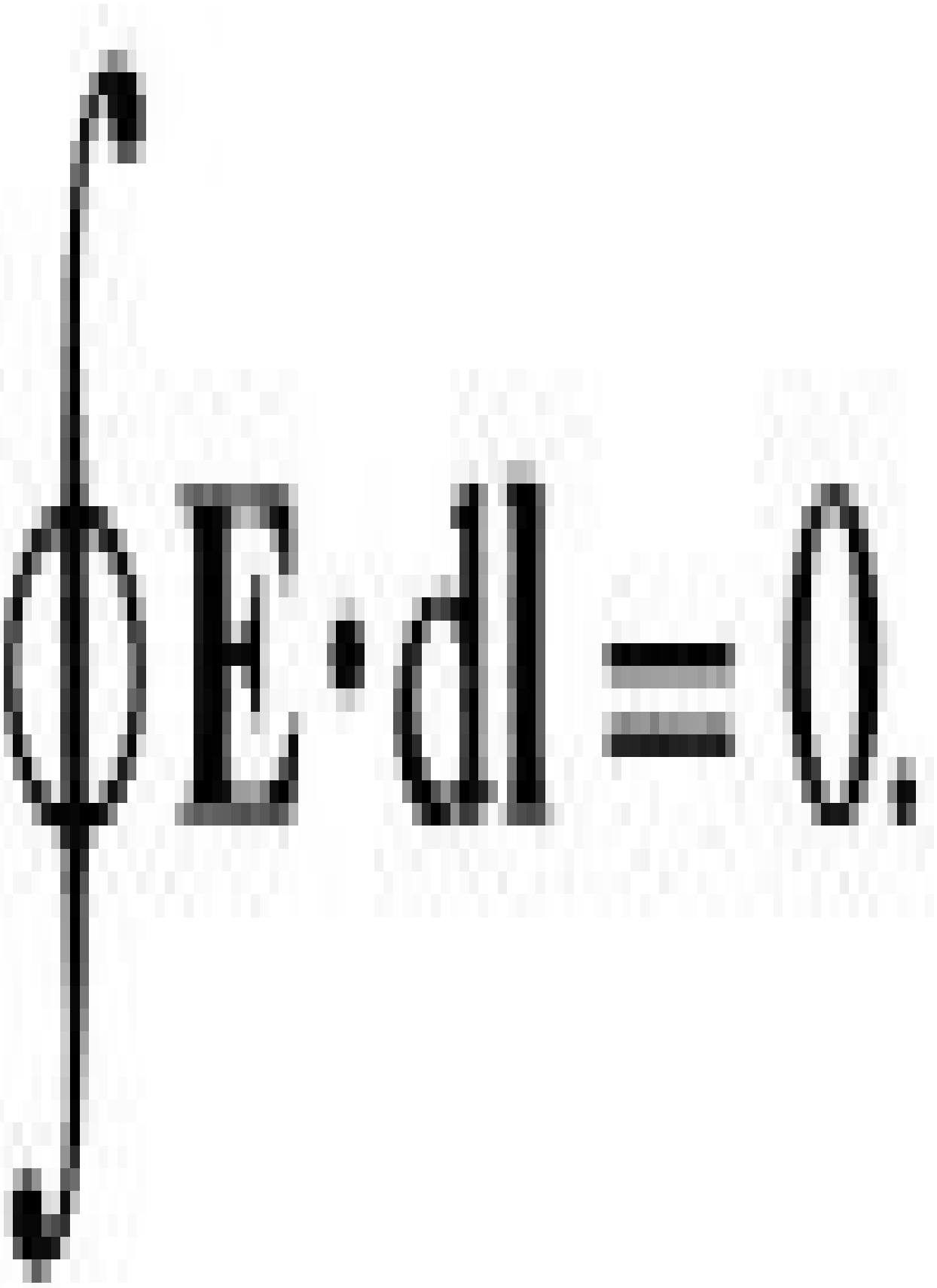


Figure 4.13 shows the variation of this line integral as we proceed along the path described above from left to right, along which the elementary vectors dl also point to the right. If the path is now closed from right to left at large distances, where the field, which is inversely proportional to the cube of the distance, tends to zero, there is no contribution to the integral from this extra part and over the complete path



It can be seen that the reversal of the field between the layers of opposite sign charge is responsible for this result.

Now let us look at the somewhat analogous situation in which a magnetic field is produced by a circular coil carrying a current I. This field is equivalent to that given by a sheet of magnetic dipoles, and has the same spatial variation, other than in the immediate vicinity of the coil, as that of the electric field of an electric dipole sheet. The magnetic field can be obtained from the electric field by replacing the electric dipole moment per unit area Np by the magnetic dipole moment per unit area I, and by replacing the factor $l/4\pi\varepsilon_0$ by $\mu_0/4\pi$, as Equations (1.34) and (4.29) indicate. The line integral of the field B over the path from a large distance on the left to very near the dipole sheet on the left has the same form as that for the electric field, and just to the left of the sheet has the value $+\mu_0 I/2$ compared with $+Np/2\varepsilon_0$ for the electric field. As the path of integration goes through the current loop we now have a big difference between the B and E fields. The lines of the field B are continuous. There are no magnetic ‘charges’ and the field lines do not reverse direction. The curve showing how the line integral varies with distance is thus continuous through the surface of the loop. On the right-hand side of the loop the curve again has the same shape as the curve for the line integral of the electric field, since the field distributions are the same. The resulting curve versus distance is shown in Figure 4.14. If we complete a closed path of integration at very large distances, there is no

Figure 4.13. The variation of the line integral of the electrostatic field E due to a sheet of electric dipoles. The curve shown corresponds to the path from a point a far to the left on axis, to a point b far to the right on axis. The schematic figure below the curve shows, on an expanded scale for clarity, the sheets of charge which constitute the dipole sheet and the associated electric field lines. The reversal of the electric field between the layers of charge results in zero value for the line integral around the complete path from a to b closed at great distance.

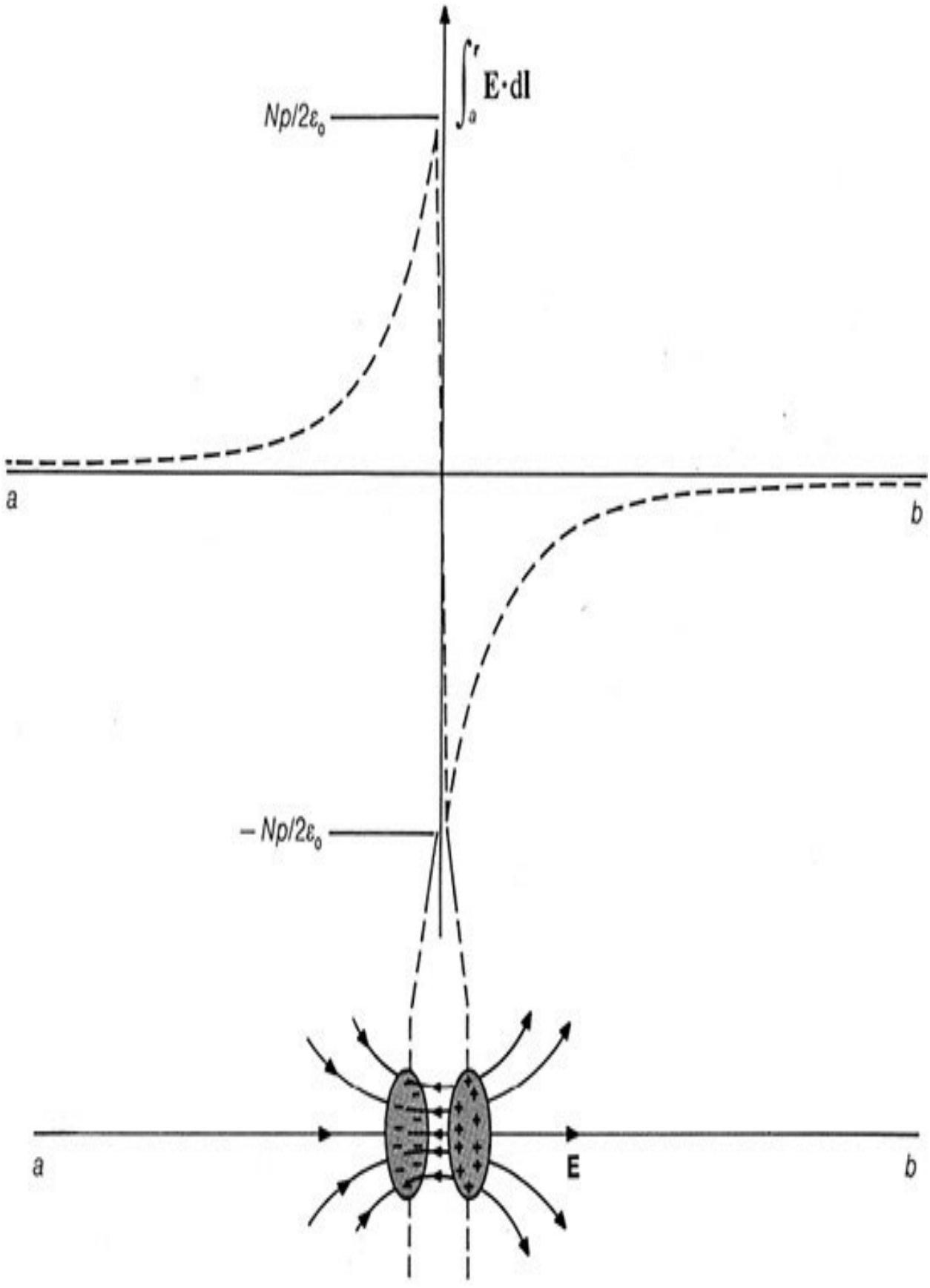
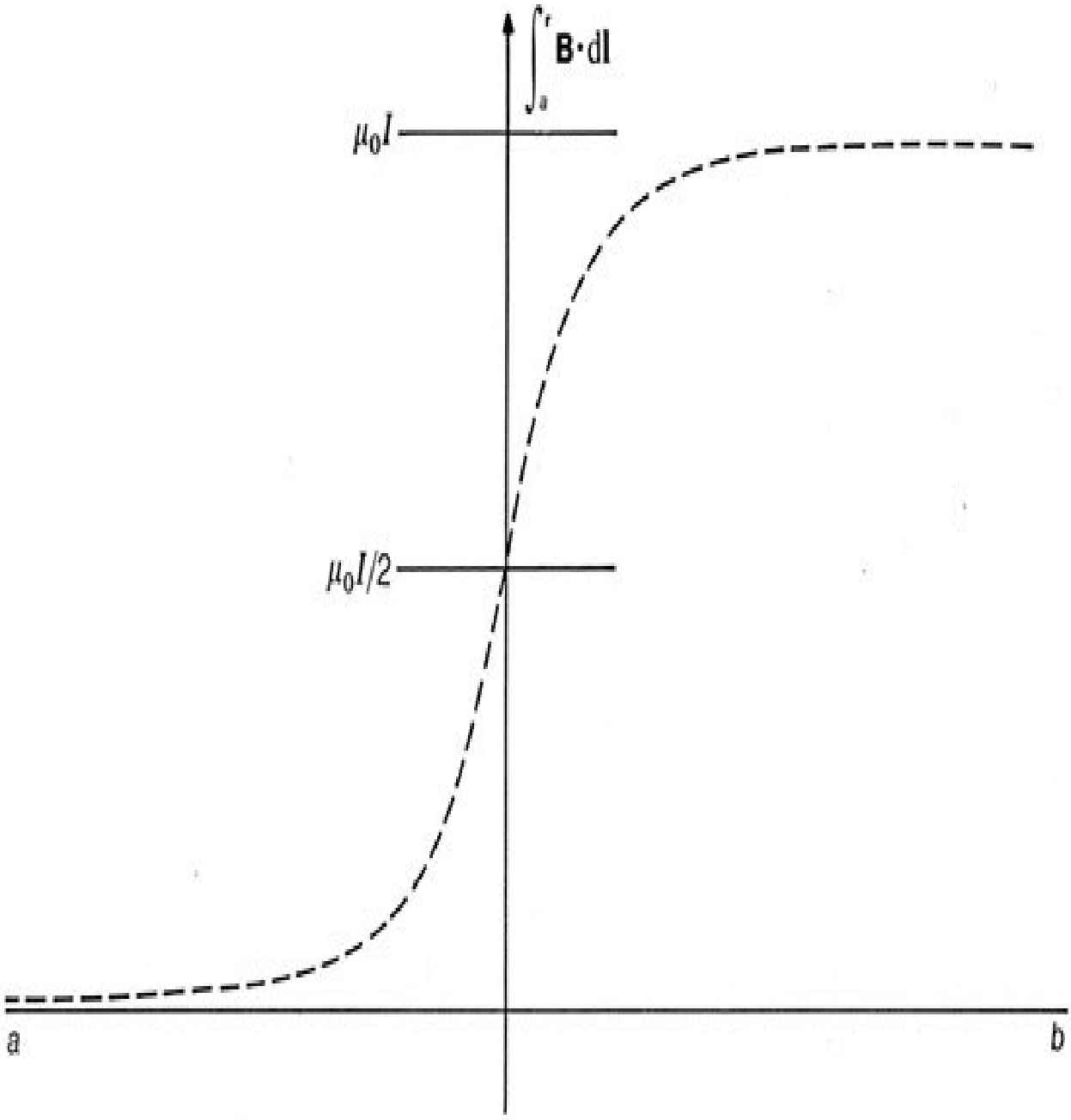
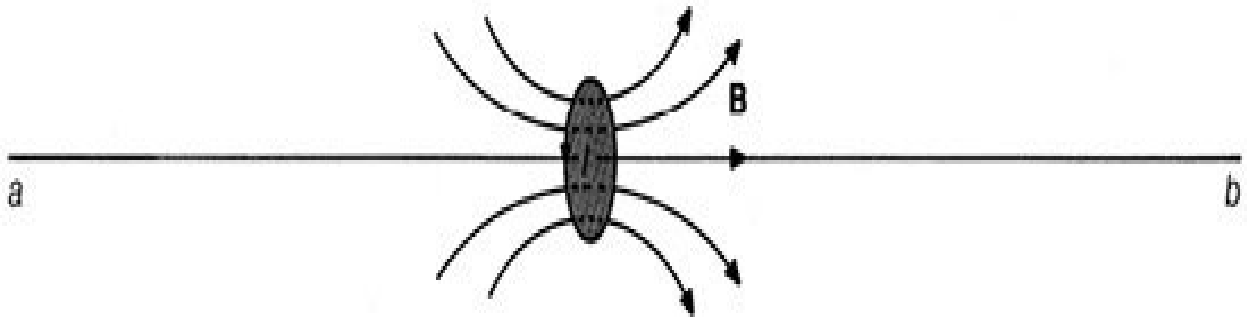
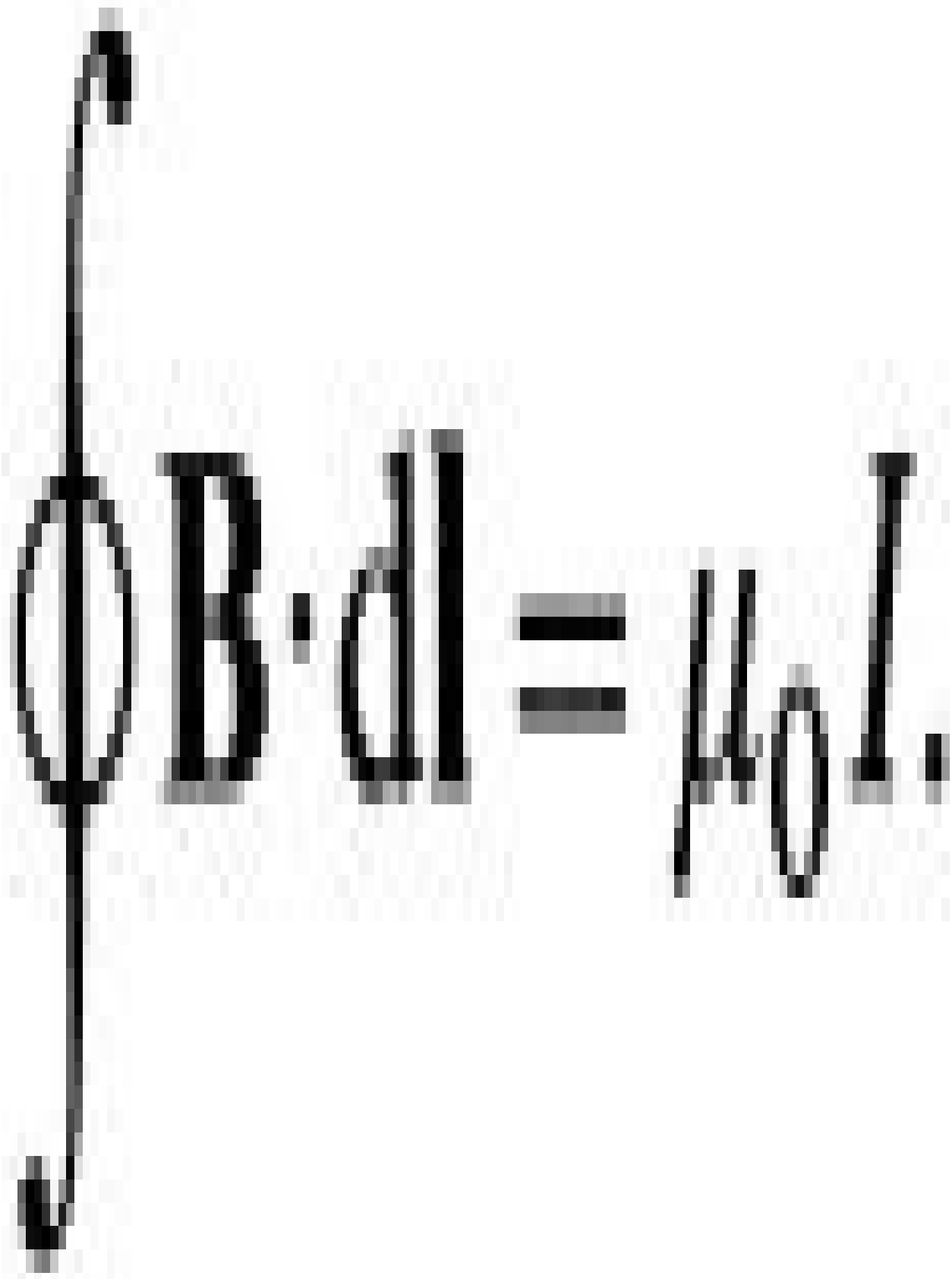

$\int_a^r \mathbf{E}\cdot d\mathbf{l}$
$Np/2\varepsilon_0$
a
b
$-Np/2\varepsilon_0$
a
E
b

Figure 4.14. The variation of the line integral of the magnetic field B due to a current in the loop of wire shown underneath. The line integral around the complete path from a point a far to the left on axis to a point b far to the right is non-zero because the lines of the field B are continuous through the loop.

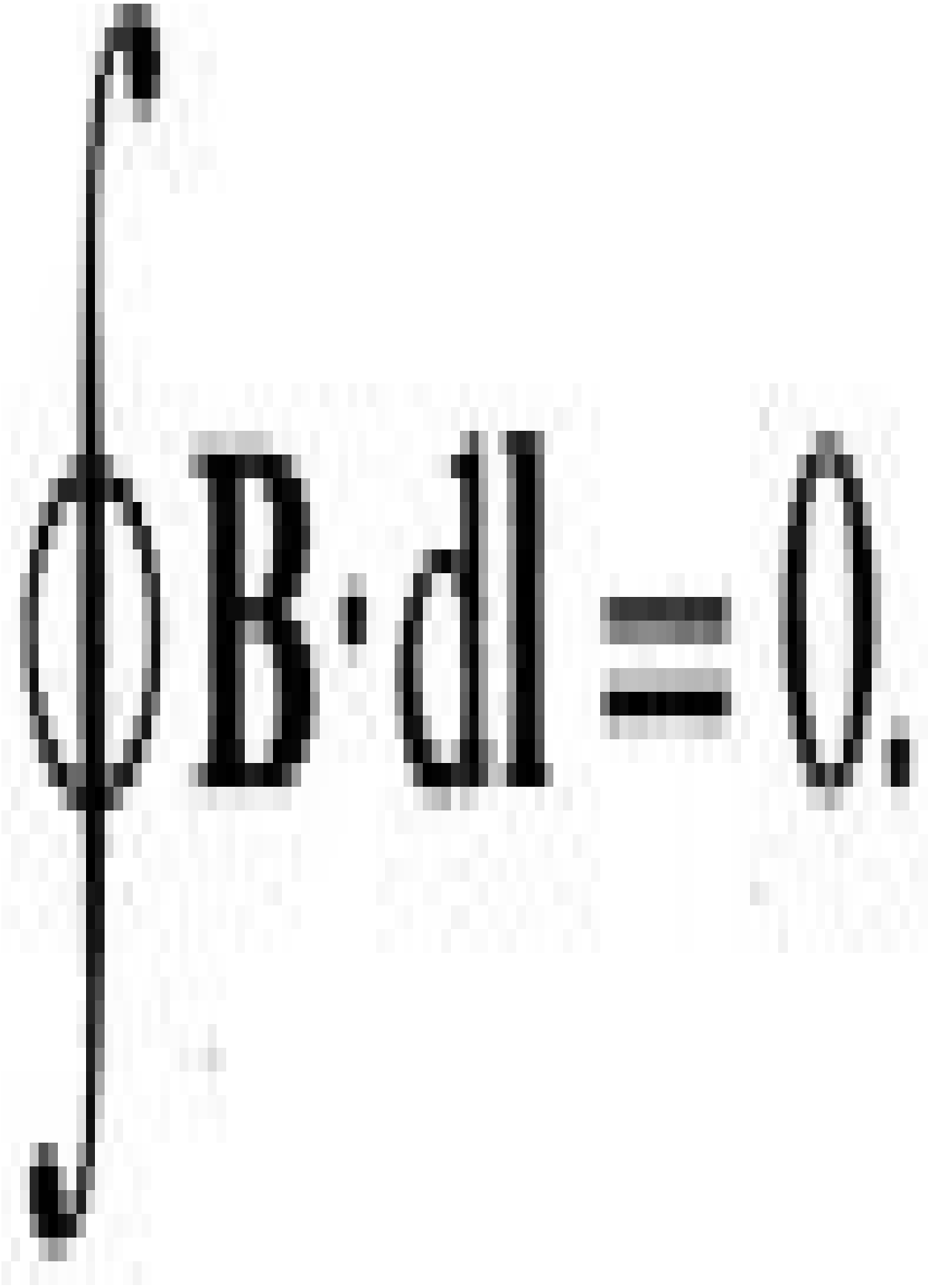

$\int_a^b \mathbf{B} \cdot d\mathbf{l}$
$\mu_0 I$
$\mu_0 I/2$
a
b


B
I
a
b

contribution to the integral from this part, and



This expression remains true for any closed path that threads the current loop and not just the one on axis that we considered. If we evaluate the line integral of the field B over a plosed path which does not go through the current loop, the variation of the magnetic field over the whole path is exactly the same as the variation of the electrostatic field of the electric dipole sheet, and



These results can be summarized in Ampère's law, which states that the line integral of the field B around any closed path is equal to constant μ0 times the current flowing through the area enclosed by the path,

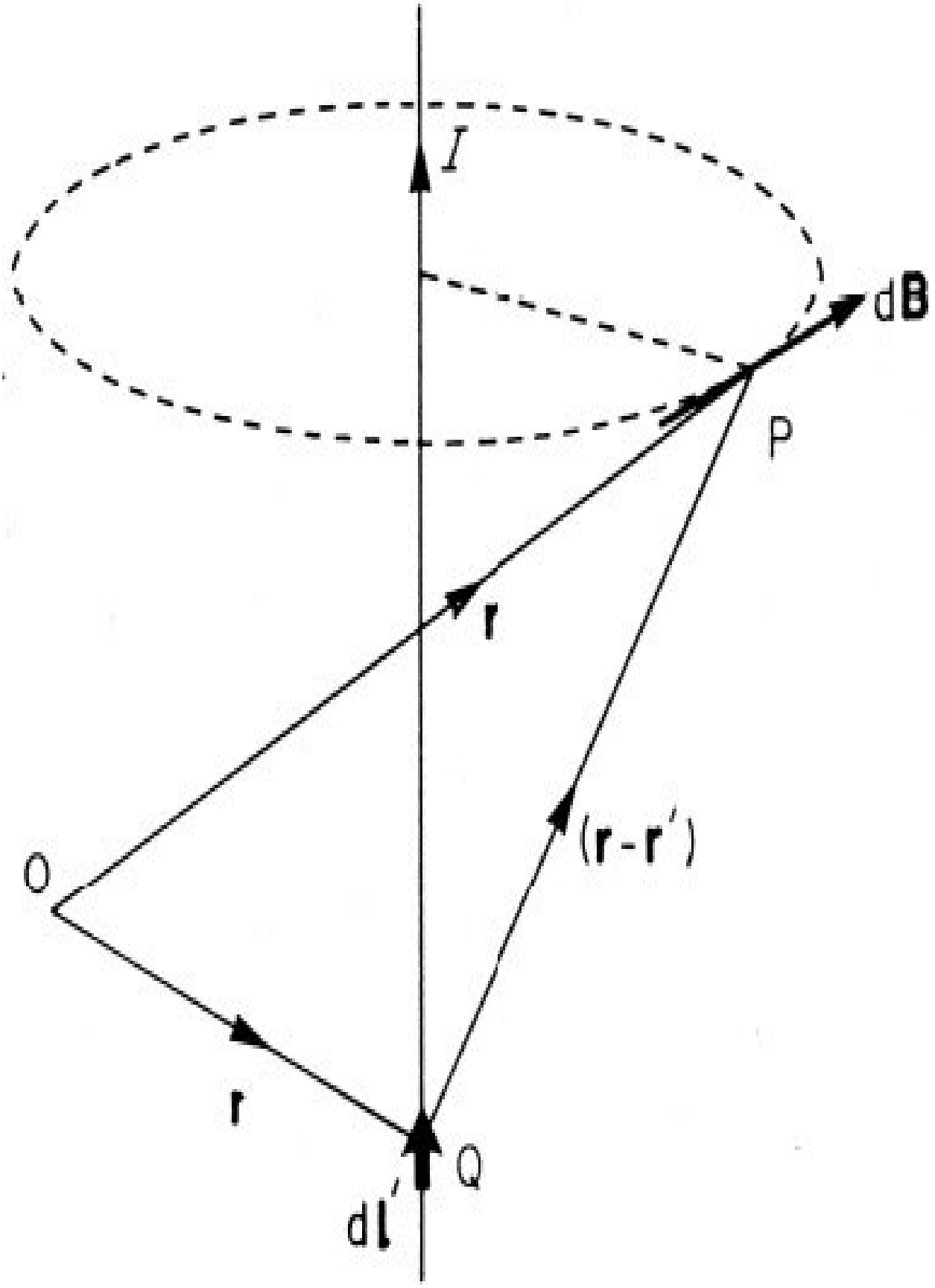
(4.31)

The sense of traversal of the loop when the line integral is performed is determined by the direction of the current together with the right-hand screw rule.

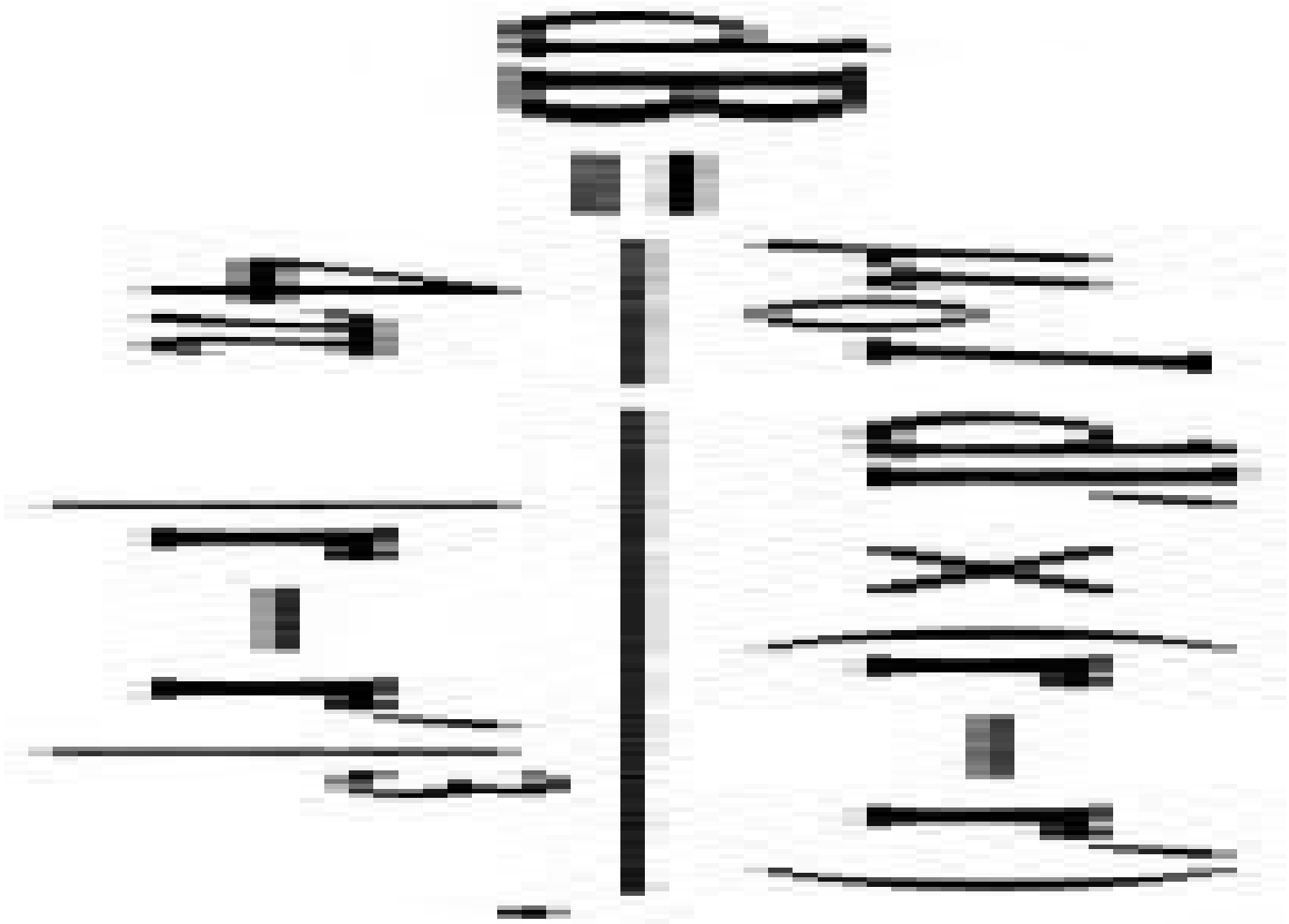
4.4.2 The Biot–Savart law

The Biot–Savart law gives an expression for the magnetic field that arises when currents flow in thin wires in free space, or, for most practical purposes, in air. This expression is analogous to Equation (1.10) for the electrostatic field of a distribution of charges, and gives the magnetic field as the sum of contributions from the infinite number of elementary lengths which comprise the circuits. The Biot–Savart law was deduced from measurements on the fields due to currents in wires made by Ampère, Biot and Savart in the years around 1820. Ampère's law can be deduced from the Biot–Savart law and can thus be regarded as distilled from experimental observations in the same way as Gauss' law was the result of observations on the forces between charges.

Figure 4.15. The field at the point P is made up of contributions dB from elements like the one shown at Q.


I
dB
P
r
(r - r′)
O
r
Q
dl′

We will introduce the Biot–Savart law by using it to determine the field due to a current in a complete circuit s, part of which is the straight section shown in Figure 4.15. The field at the point P (position vector r with respect to the origin O) can be considered to be made up of contributions dB from elements of length dl′ like the one at the point Q. The magnitude of the contribution from each element is proportional to the current I in the wire, and inversely proportional to the square of the distance of the element from the point P. Let the position vector of the point Q be r′. The direction of the contribution dB from the element at Q is perpendicular to the vector dl′, which is in the direction of the current in the wire at the point Q, and perpendicular to the vector (r – r′) which is in the direction of the line joining Q to P. Combining all the properties listed above we can write the contribution dB as



The total field at the point P is obtained by integrating around the complete circuit s,

(4.32)

This is the Biot–Savart law.

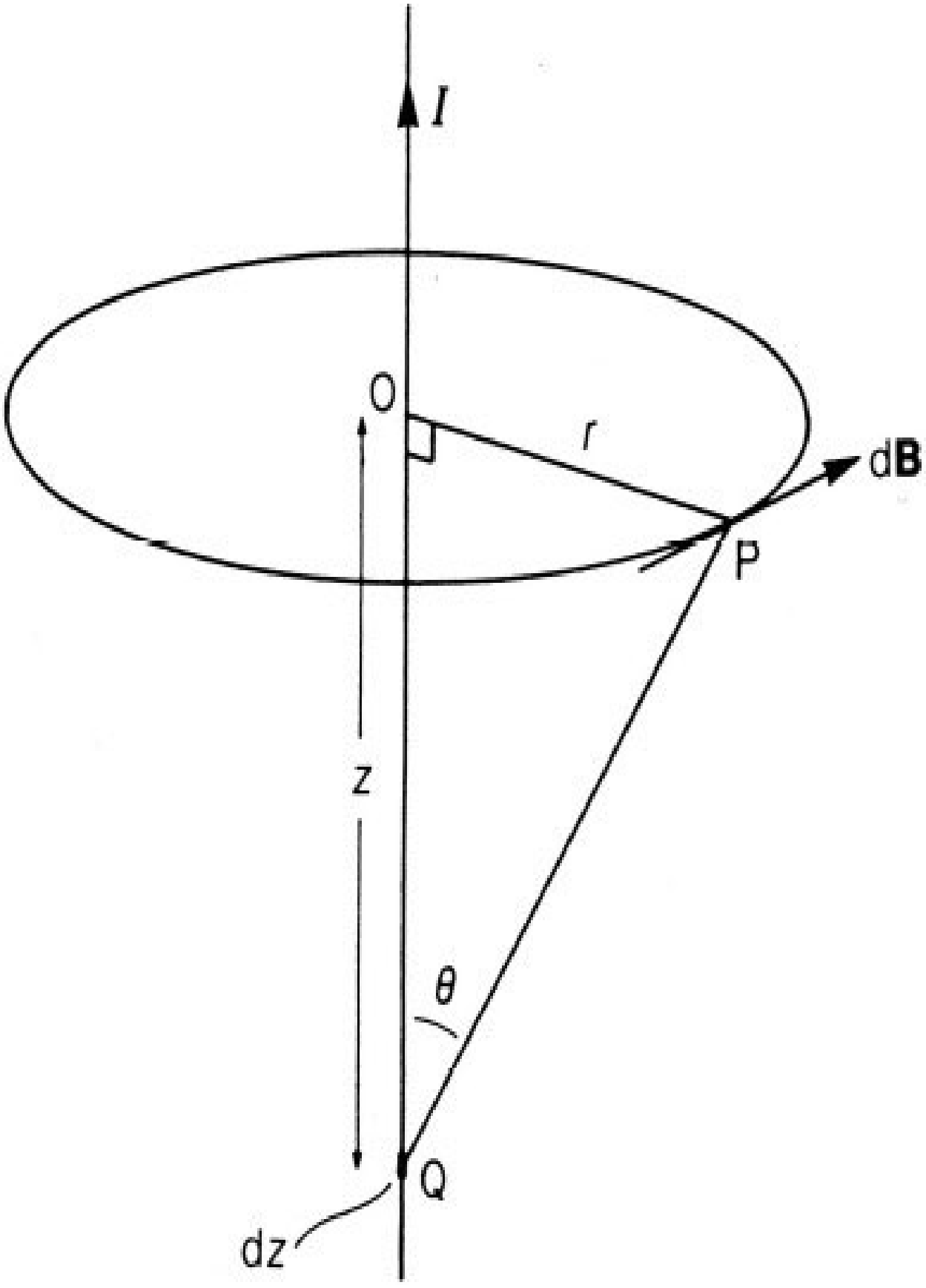
Let us work out the field at a point P when the circuit has a long straight section. If the distance of the point P from the wire is small compared to the length of the straight portion and the circuit is closed far from P the field can be considered with little error to be that arising from a current in an infinitely long straight wire. It is simplest to choose the origin of coordinates where the perpendicular from P hits the wire, as shown in Figure 4.16. Let the distance OP be r, and the distance OQ be z. The variable r′ in Equation (4.32) now becomes ezz, where ez is a unit vector in the direction of the current in the wire, and the variable dl′ becomes ezdz. The vector dl′ × (r – r′) has magnitude QP dz sin θ, where θ is the angle between the lines OQ and QP, and no matter where Q lies on the wire always points in the same direction. It is tangential to the circle and in a direction related to the direction of the current by the right-hand screw rule. The magnitude of the resultant field is given by

(4.33)

(4.34)

Figure 4.16. The calculation of the field of a long straight wire carrying a

current.


I
O
r
dB
P
z
θ
Q
dz

4.4.3 Examples of the calculation of magnetic fields

Ampèere’s law enables us to determine the magnetic fields of certain current distributions whose symmetries ensure that the field lines have a simple pattern. The following examples (i) and (ii) illustrate the use of Ampère’s law and of other ideas developed in previous sections, while example (iii) is another application of the Biot–Savart law.

(i) The field inside a long solenoid

Consider a long solenoid with N turns of wire per unit length and carrying a current I, as shown in Figure 4.17. Ampère’s law requires that the line integral of the field B over the closed path ABCD shown in the figure be given by

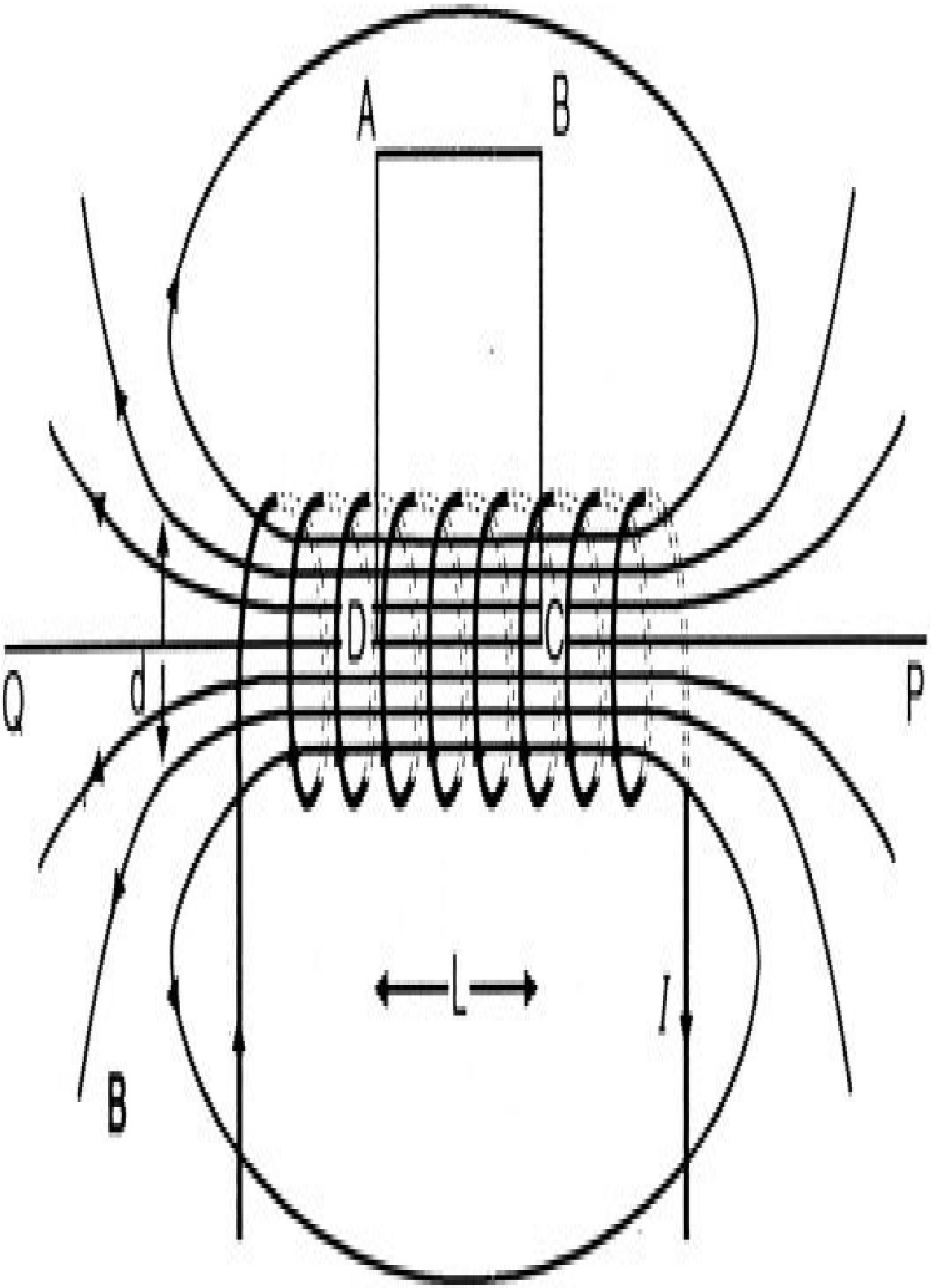
(4.35)

The term on the right-hand side is μ0 times the total current that flows through the area traced out by the path chosen for the line integral. Here the area is threaded by NL turns of wire each carrying a current I. The path has to be traversed in the direction from P to Q in order that the sign of the term NLI be positive.

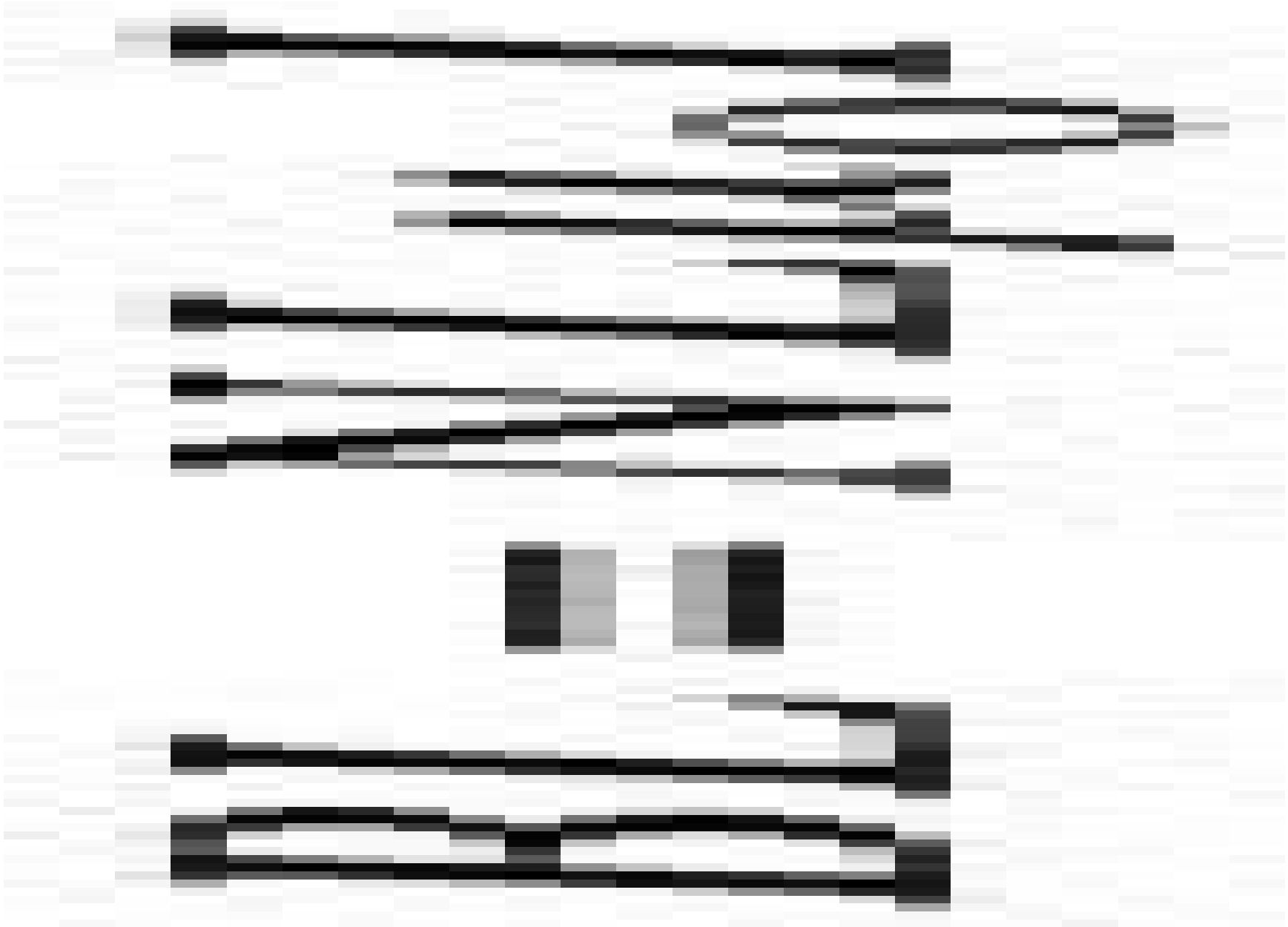
The resultant field is the sum of the fields from all the coils of the solenoid, each of which reinforce the field in the interior, but which tend to cancel each other outside. This can be appreciated by regarding the long straight solenoid to be

part of a much longer solenoid bent to form a toroid of very great diameter. The field outside this toroidal solenoid is by symmetry the same at all points on a circle with centre on the axis of the toroid. The line integral of the field B around such a complete circle is equal to the current through the area defined by the circle. This current is zero and hence the field at all points on the circle is zero. This situation remains the same for all circles with centre on the axis of the toroid and hence at all points outside the toroid the field B is zero. Similar reasoning leads to the conclusion that the field is constant and non-zero at all points inside the toroid and lies along the solenoid axis in the direction given by the right-hand rule.

Figure 4.17. The field of a long solenoid carrying a current. The field is very nearly uniform inside the solenoid.


A
B
D
C
Q
P
L
I
B

The actual field in and around the long straight solenoid has a shape given by the field lines shown in Figure 4.17. Let the field in the middle of the solenoid have magnitude B. Equation (4.35) then becomes



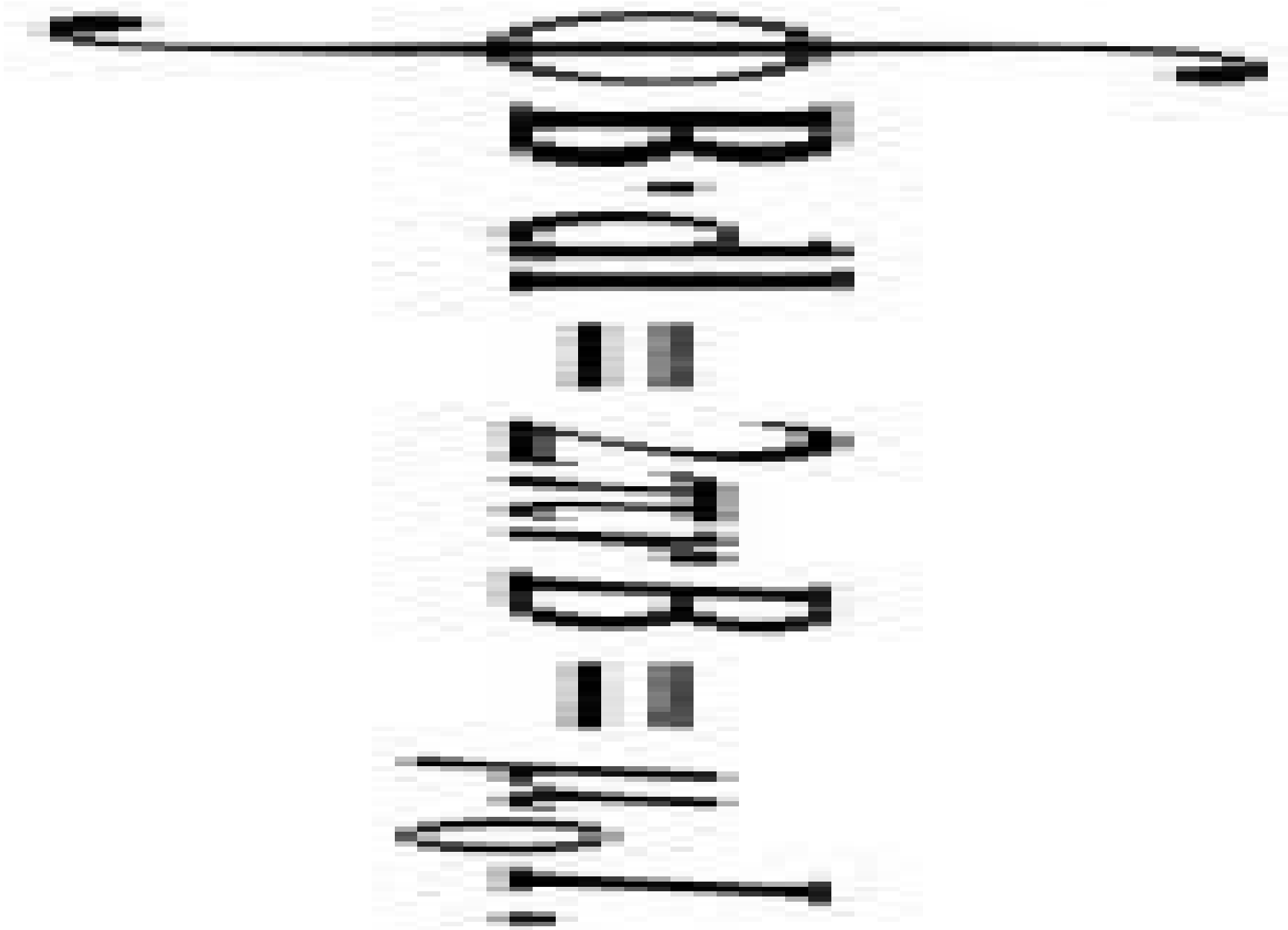
or

(4.36)

This equation predicts for example that one ampere flowing in a solenoid which has 20 turns per cm produces a field inside equal to about $(20 \times 10^2) \times (4\pi \times 10^{-7}) \times 1.0$; about 2.5×10^{-3} T. The approximations made above become better as the ratio of the length l to the diameter d of the solenoid increases. Equation (4.36) gives the field in the middle of a solenoid to better than 1 % if $l/d > 10$.

(ii) The field due to a current in a long straight wire

Another simple example of the application of Ampère's law is that of the field due to a current I in a long straight wire, already considered using the Biot–Savart law in section 4.4.2. In this situation the field has cylindrical symmetry; its magnitude must be the same at all points whose perpendicular distances from the wire are equal. Figure 4.18 shows a circle of radius r centred on a point O on the wire, and lying in a plane perpendicular to the wire. The field at a point on this circle has constant magnitude B, lies along the tangent to the circle at the point and has a direction given by the right-hand screw rule. The line integral of the field B around the circle is given by Ampère's law,

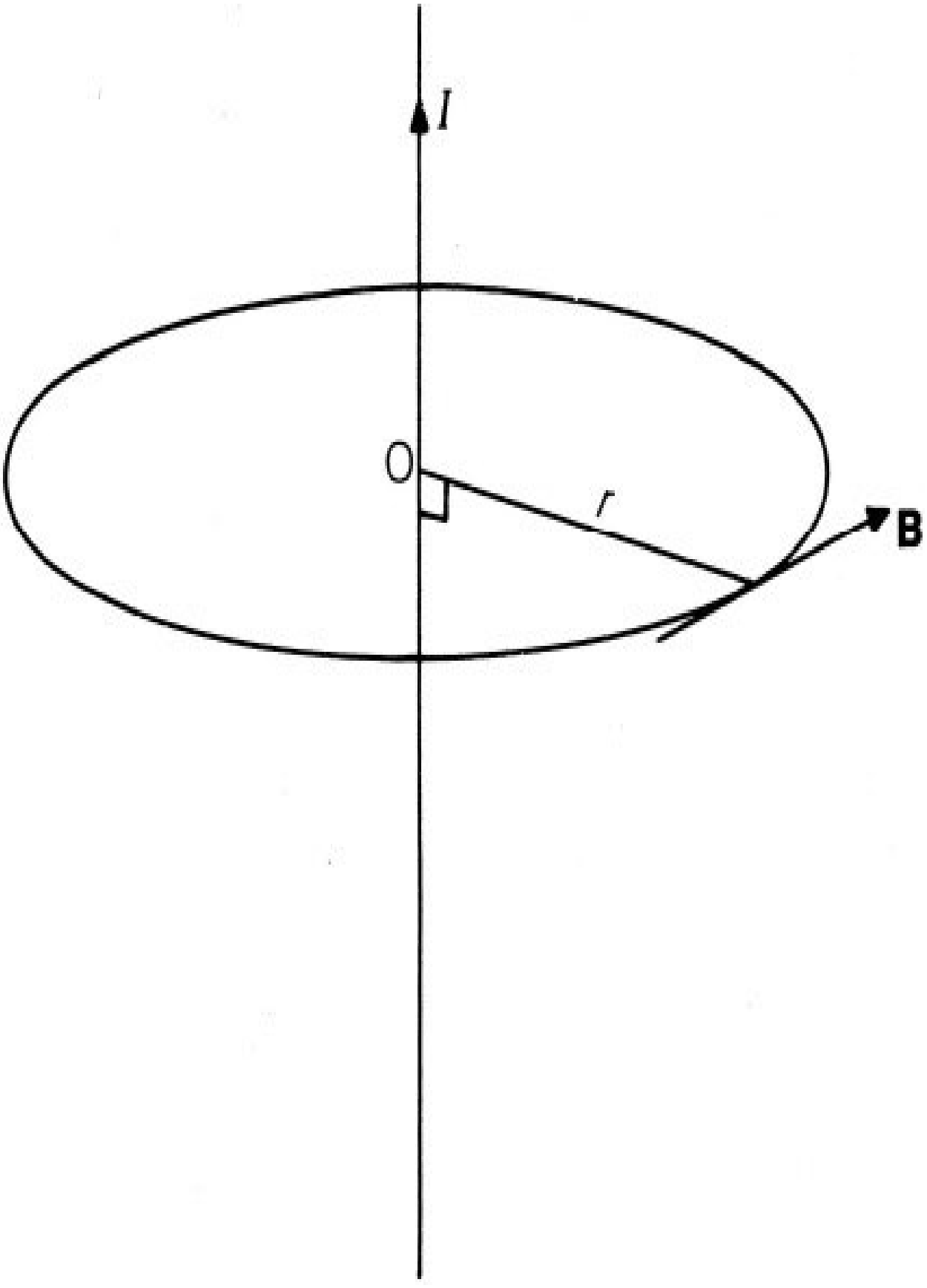


Hence

(4.37)

A current of one ampere flowing in a long wire produces a field of 0.2 gauss at a distance of one cm from the wire, a field comparable with the Earth’s magnetic field at the surface of the Earth.

Figure 4.18. The field of a long straight wire carrying a current has cylindrical symmetry. It looks the same at all points on the circle of radius r. Hence it has constant magnitude on the circle and is in a direction tangential to the circle at all points on the circumference, since the line integral of the field around the circle must equal μ0I.


I
O
r
B

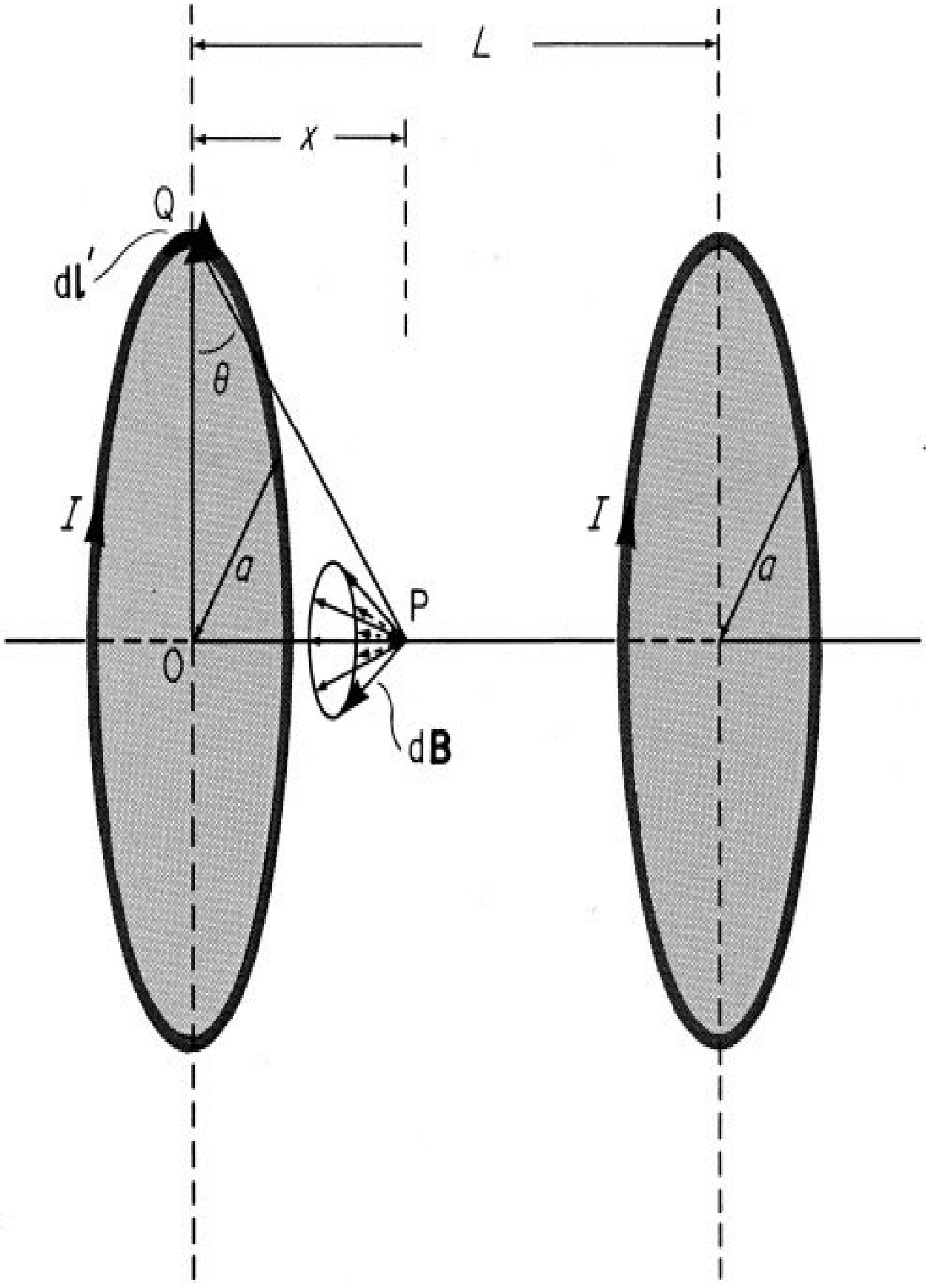
(iii) The field due to a current in coaxial coils

As another example of the use of the Biot–Savart law, Equation (4.32), we will calculate the field on the common axis of the two coils shown in Figure 4.19. Two parallel, coaxial coils arranged like this are often used to provide a field that cancels out unwanted fields, like the Earth's field, or fields from iron building supports, etc. Two coils are used because near the axis they can provide a field which is very nearly uniform over an appreciable volume, as we shall see. Let us calculate the field at the point P distant x from the centre O of the left-hand coil. We take the point O to be the origin of coordinates. Consider the contribution dB to the field of the left-hand coil from an element dl′ at the top point Q of the coil, as shown. The vector dB is in the plane OQP and makes an angle θ with the axis, the same angle as that between the lines OQ and QP. Contributions from successive elements obtained by proceeding around the coil all make the same angle θ with the axis and lie in a cone of half angle θ. The result of the vector integral 4.32 for the resultant field can be seen to be in the direction from P to O along the axis. Its magnitude is

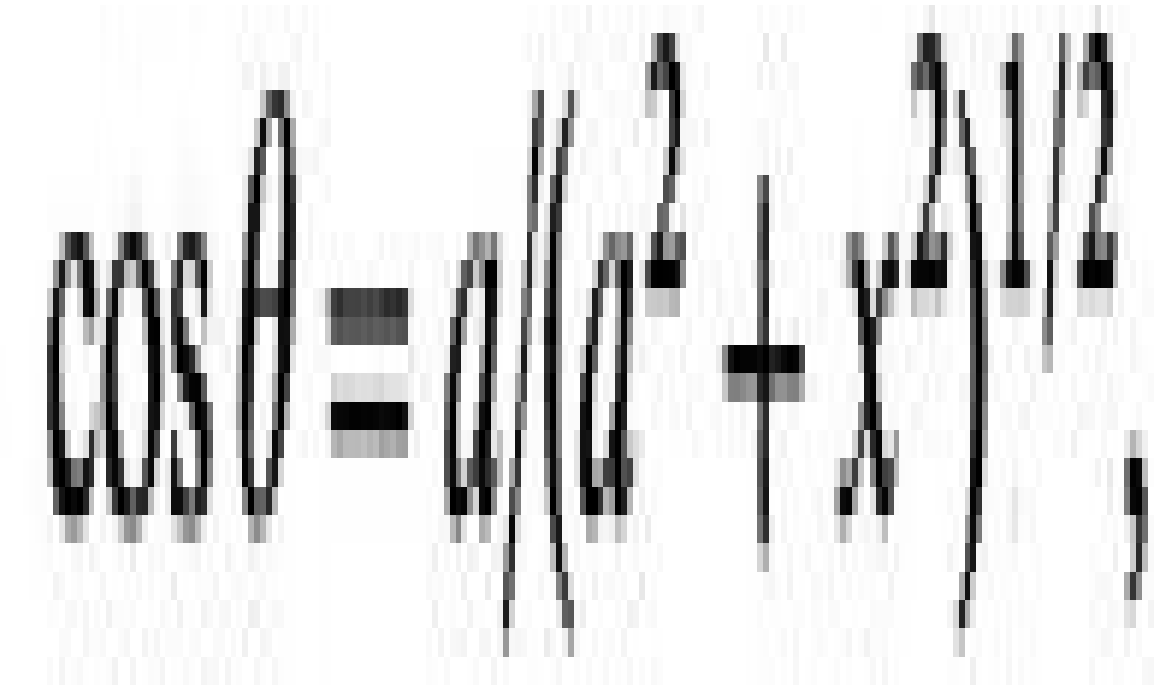
(4.38)

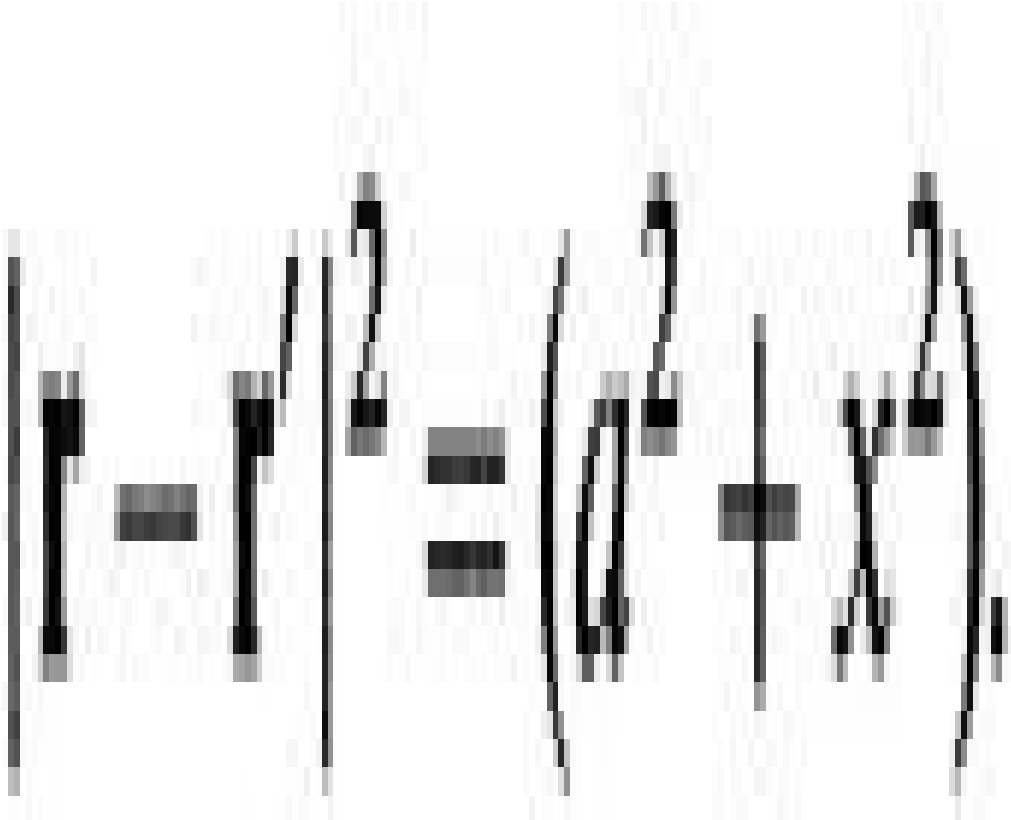
The path of the integral is around the circular coil and at all points on the path

Figure 4.19. The field due to a current in two parallel coaxial circular coils.


L
x
Q
dl'
θ
I
a
P
O
dB
I
a

the terms $\cos\theta$ and $|r - r'|^2$ are constant and given by





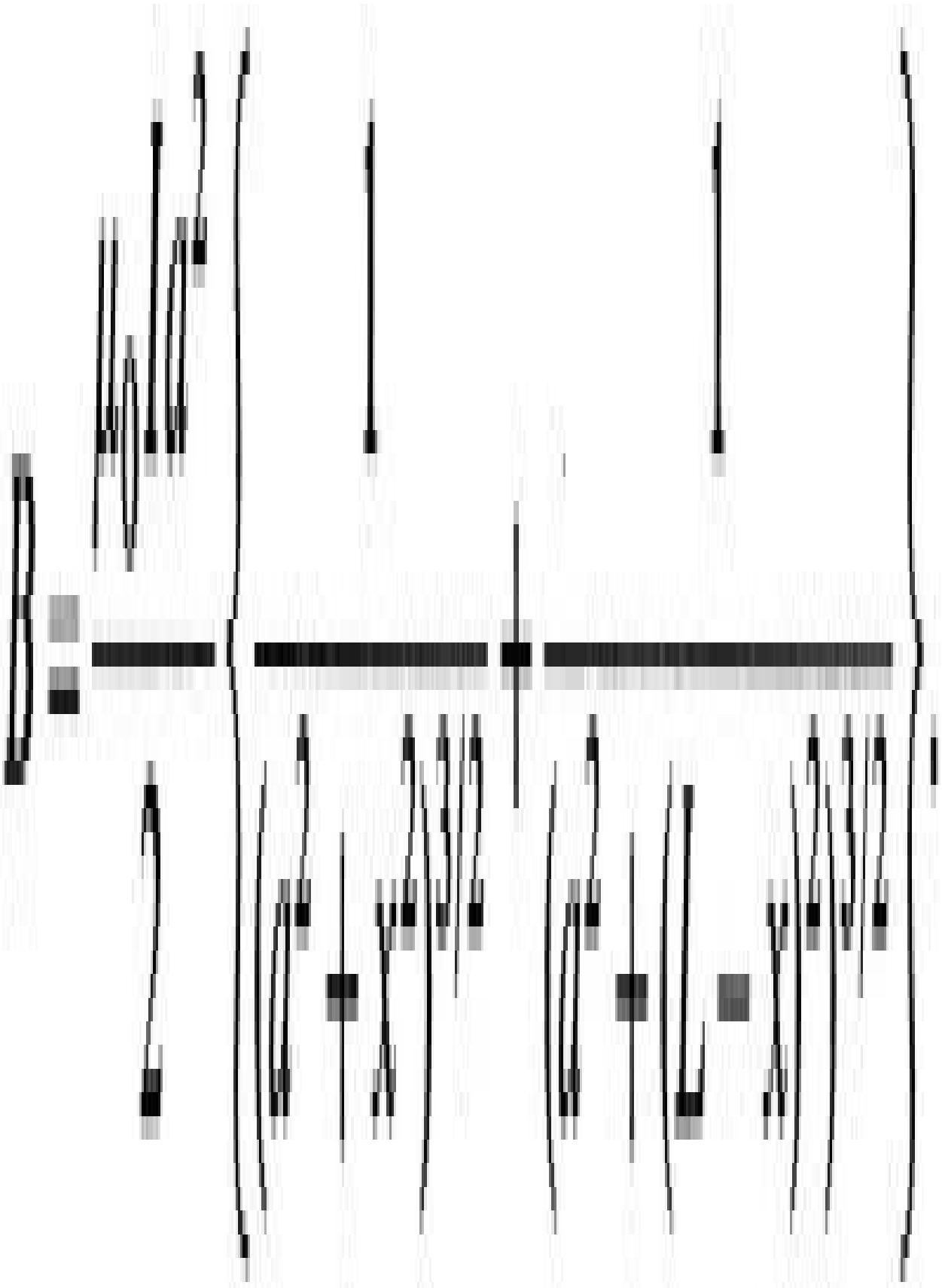
Substitution in Equation (4.38) gives

(4.39)

(4.40)

The field at P due to the second coil is in the same direction, and its magnitude is simply obtained by replacing the variable x in the above equation by (L – x). Hence the resultant field is

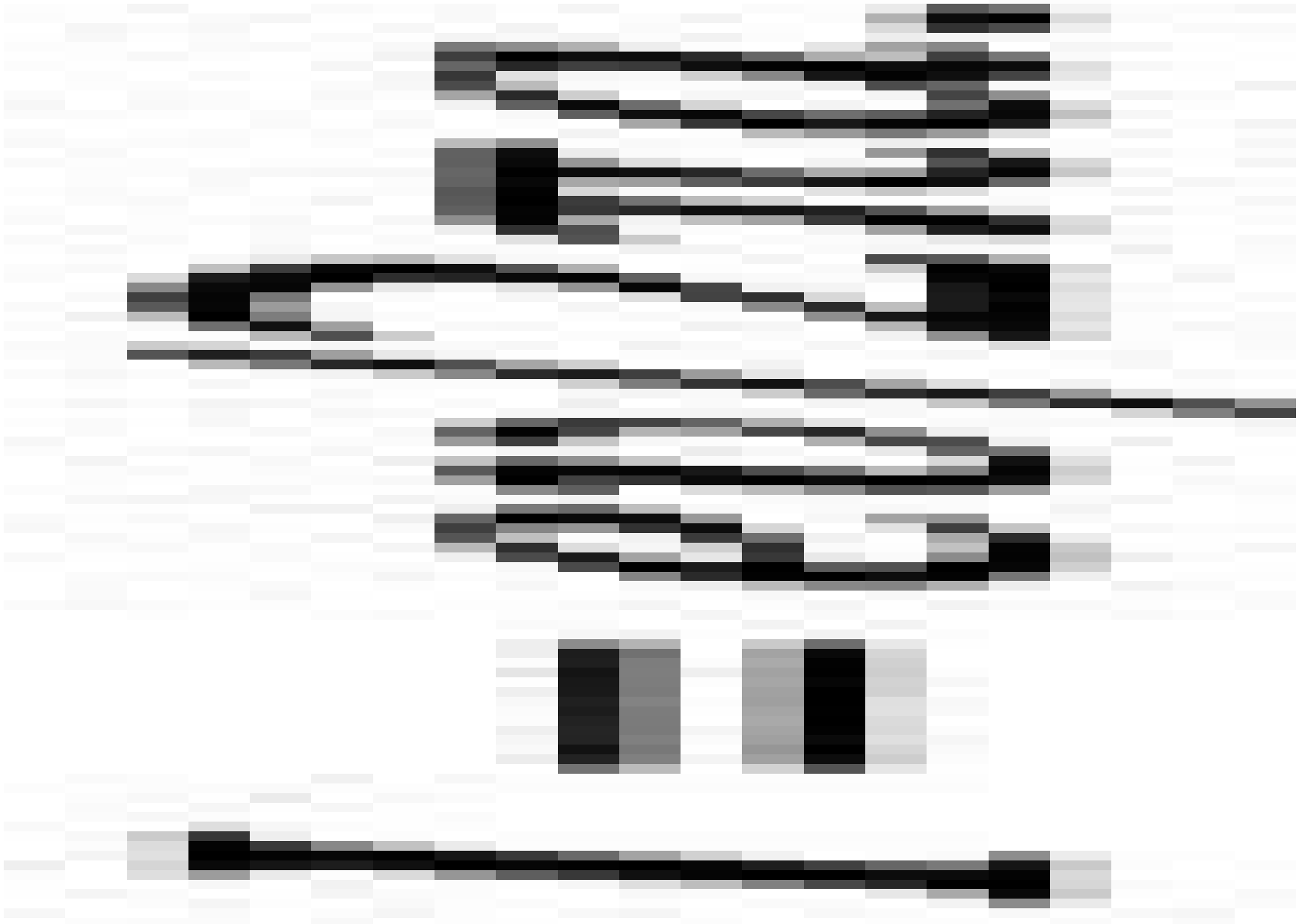
(4.41)



The field near to any point can be expanded in a Taylor series in terms of the derivatives of the field at the point. If the first few derivatives of the field are zero the field will be very nearly uniform near that point. The first three derivatives with respect to the distance x in expression 4.41 are zero at the point half-way between the coils if the coils are separated by a distance equal to their common radius. Similar, coaxial coils, separated by a distance equal to their common radius, and carrying equal currents in the same direction, are called Helmholtz coils. Suppose that a pair of Helmholtz coils, each of radius 1 m and having 200 turns is used to nullify the effect of the Earth's magnetic field at a place where the magnitude of this field is 0.6 gauss. The current through the coils needed to produce such a field in the middle is given by Equation (4.41) (after putting L = a = 1 m, and x = 0.5m) to be about 0.3 A.

(iv) The magnetic fields inside atoms

We will use a very crude model of an atom in order to guess the order of magnitude of microscopic magnetic fields. Consider an electron moving in a circle of radius a around a nucleus with constant speed v. The circular motion of charge is equivalent to a current I around the circle, and gives rise to a magnetic field of magnitude B at the nucleus. The current I is equal to the charge that passes any point on the circle every second, i.e.



Hence, from Equation (4.40), with $x = 0$ we obtain

$$B = \mu_0 ev / 4\pi a^2$$

$$= \frac{\mu_0 e^2}{4\pi m_e a^3} (m_e v a),$$

where me is the mass of the electron, and the last equation has been written in terms of the angular momentum (meva) of the rotating electron. The reason for doing this is that the angular momentum of an atomic system is restricted to certain definite values by the laws of quantum mechanics. We are already familiar with the quantization of certain physical observables; for example electric charge is always observed in multiples of the electronic charge e. The quantum unit of angular momentum associated with orbital motion of atomic particles is $h/2\pi$ where h is Planck's constant, which has the value 6.626×10^{-34} joule second. The component in a specified direction of the orbital angular momentum of an electron in an atom must be an integral multiple of $h/2\pi$ or zero. Taking the value of (meva) to be $h/2\pi$, in order to make a rough estimate of the magnetic field, we obtain

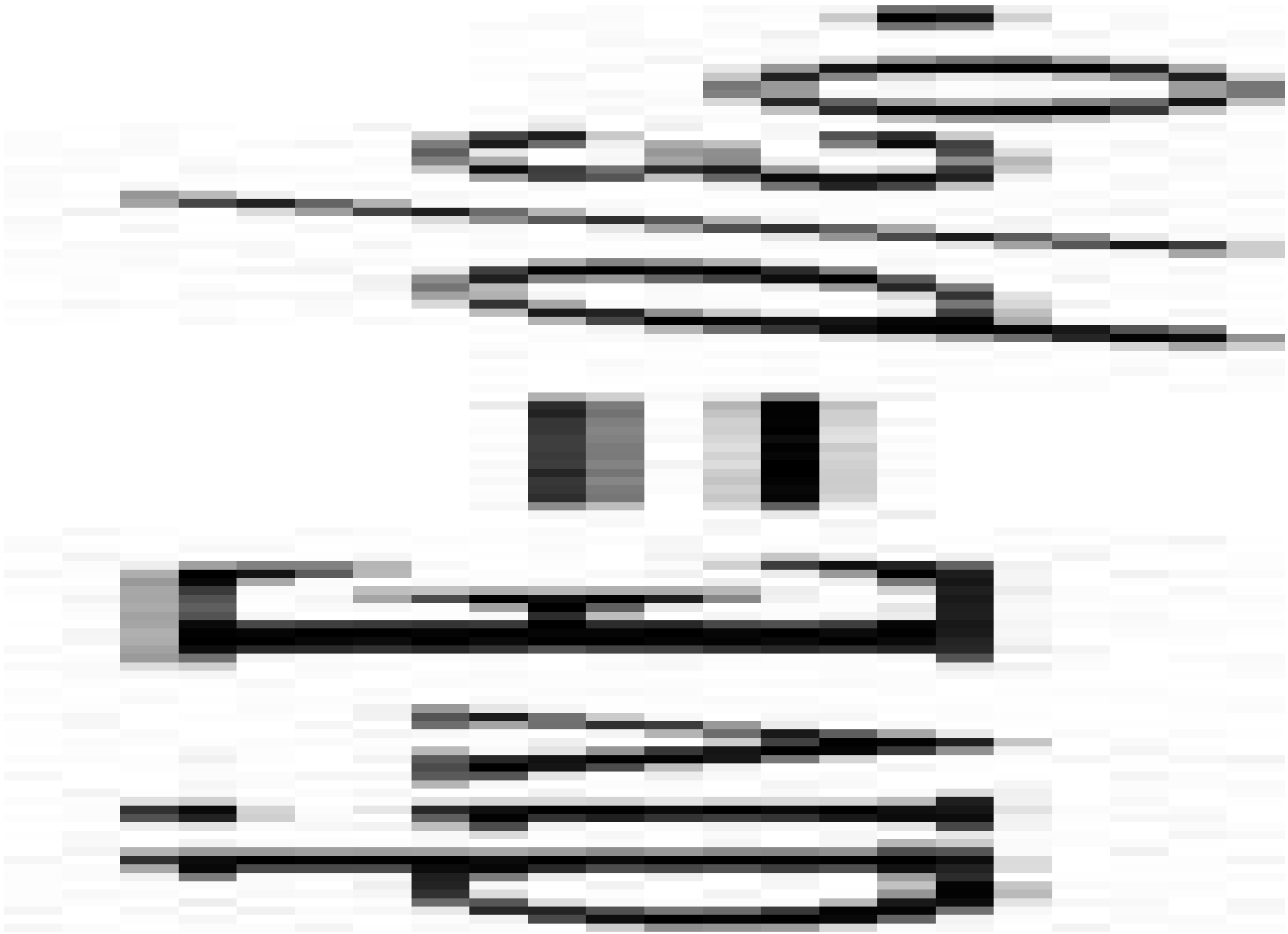
(4.42)

Substituting in this expression the value 10^{-10} m for an atomic radius we find that B is about 2T.

The crude classical model we have used is not a proper description of an atom, and in addition magnetic fields inside atoms and molecules have important contributions from intrinsic magnetic moments possessed by electrons. Nevertheless our calculation does suggest that microscopic magnetic fields are much greater than those typically encountered when currents flow in macroscopic conductors. This is in agreement with experimental observations.

4.5 THE DIFFERENTIAL FORM OF AMPÈRE'S LAW

We have already stated that Ampère's law 4.31 is as basic for the understanding of the effects that arise when steady currents flow in conductors as Gauss' law (1.14) is for the understanding of electrostatics. The next step in the study of magnetostatics is to reduce the integral form of Ampère's law to a differential equation, just as we reduced the integral form of Gauss' law to the differential equation

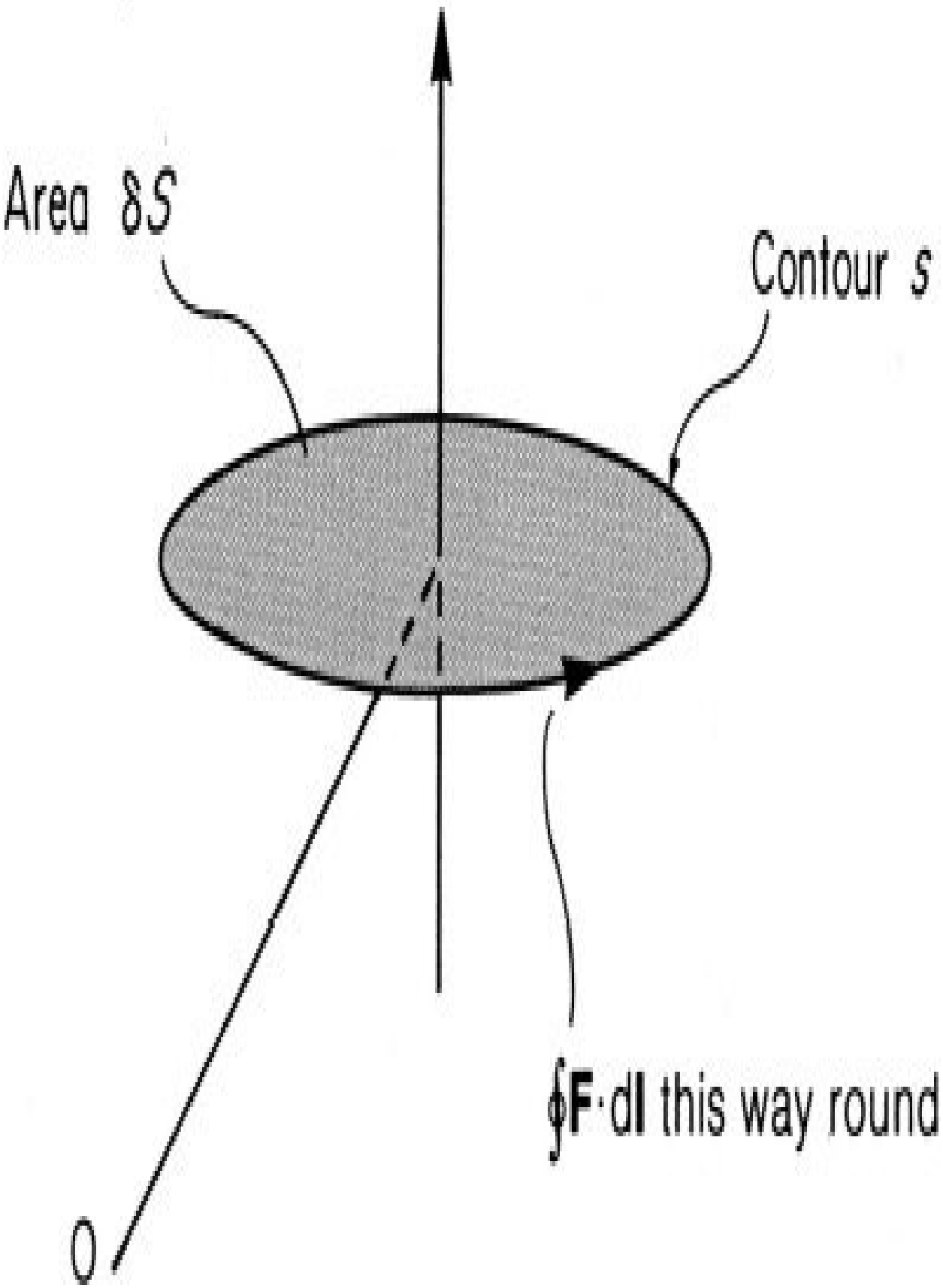


The differential form of Ampère's law is most conveniently expressed in terms of another vector differential operator called curl. We first digress in order to introduce this operator, and then show that it occurs in the differential equation we seek relating the field B to currents.

4.5.1 The operator curl

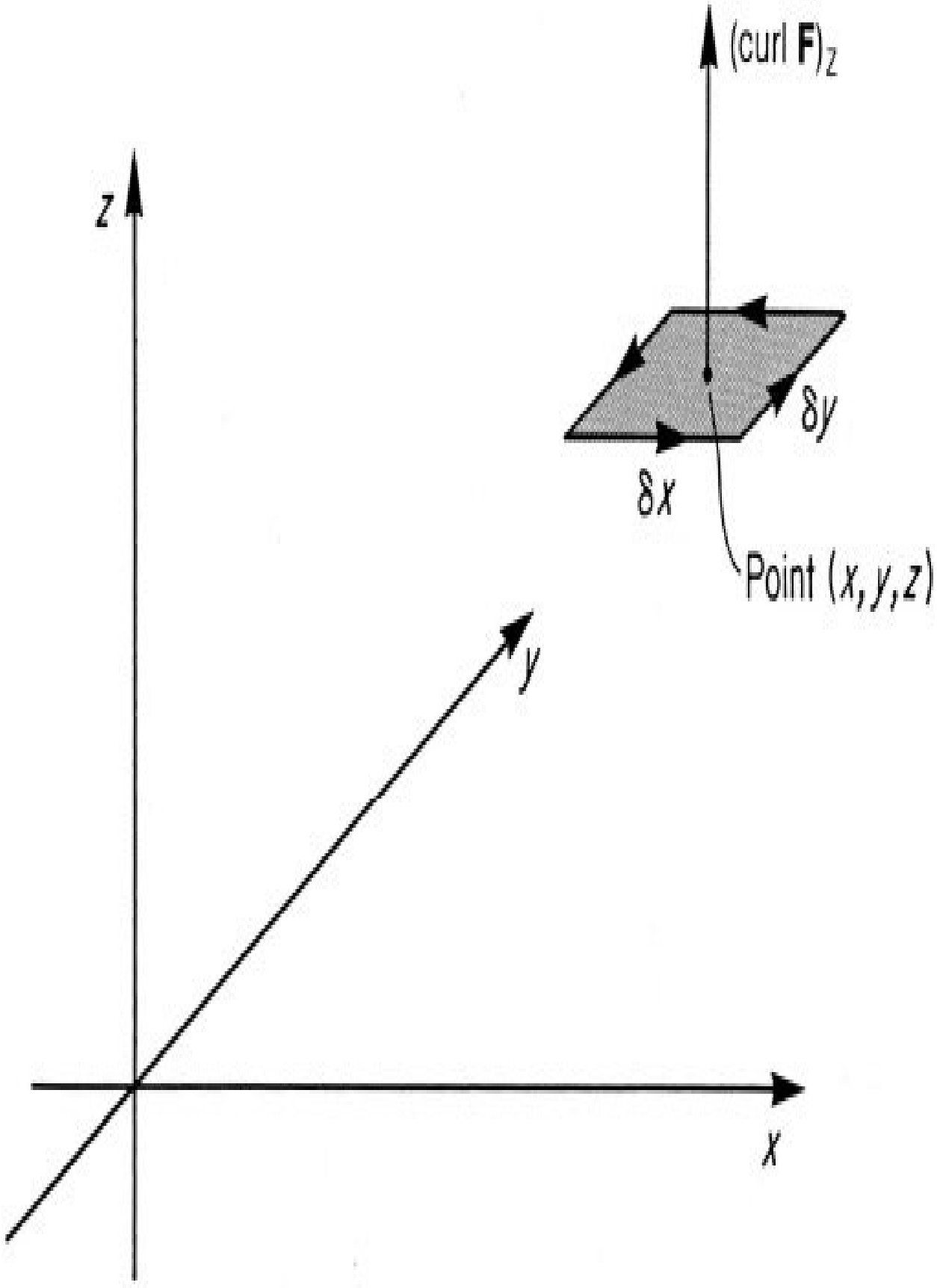
The result of performing the curl operation on any vector field F is to produce another vector field, curl F. The component of curl F in a particular direction is defined in terms of the line integral of the vector F around a closed curve drawn in a plane perpendicular to that direction.

Figure 4.20. In order to obtain the component of curl F in the direction of the arrow, the line integral $\oint F \cdot dl$ must be calculated as shown, i.e. the loop must be traversed in the direction given by the right-hand rule.


Area δS
Contour s
$\oint \mathbf{F} \cdot d\mathbf{l}$ this way round
O

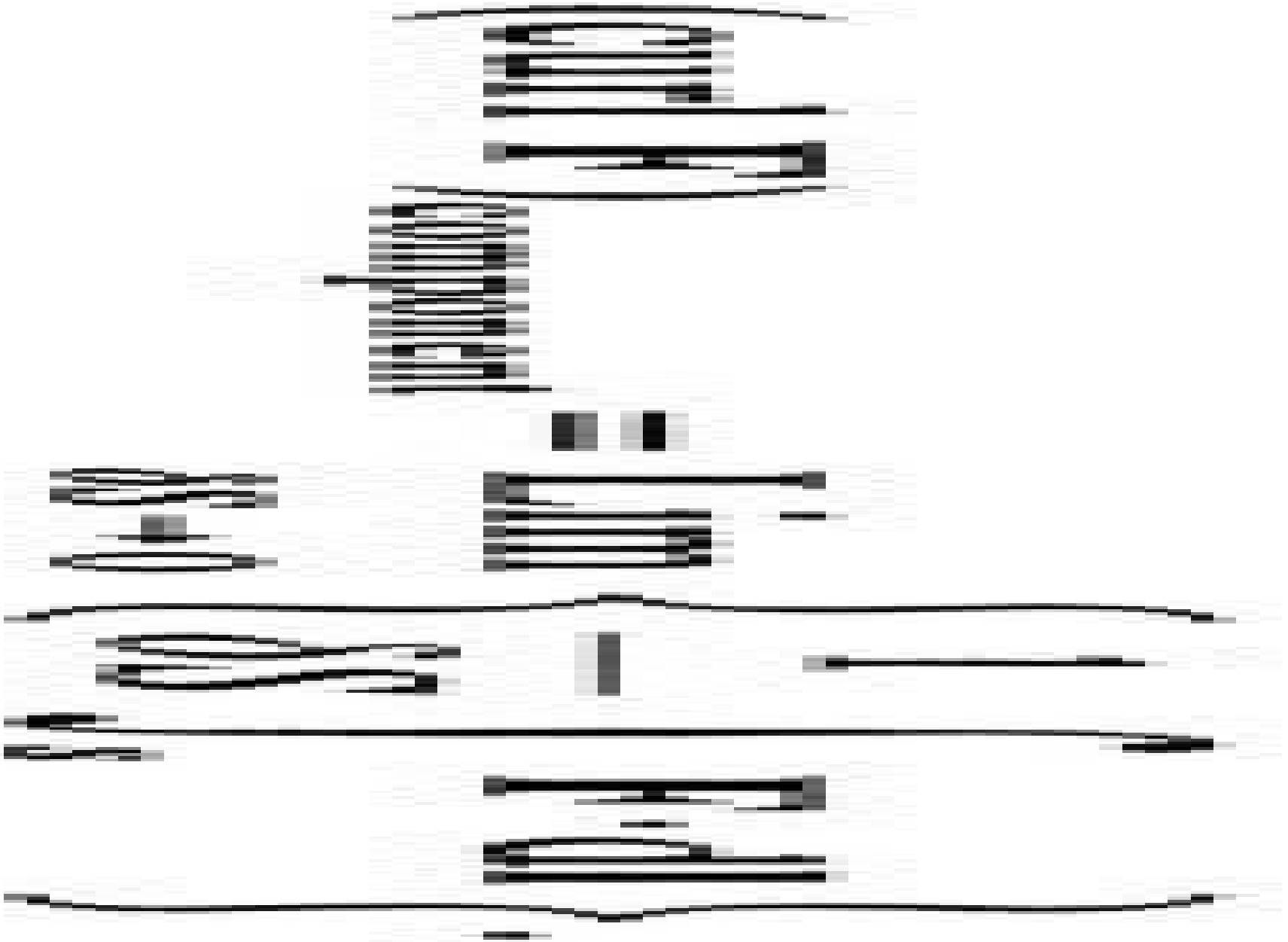
Let us examine what happens to the line integral $\oint$ F·dl around the closed path s shown in Figure 4.20, as we make the length of the path s smaller and smaller. This integral will tend to zero as the path s tends to zero. If δS is the area of the surface whose perimeter is the path s the quantity $\oint_s$ F·dl/δS may tend to a non-zero limit as δS tends to zero. The component of curl F in the direction of the arrow (given by the direction a right hand screw advances if

Figure 4.21. The evaluation of the z-component of curl F at the point (x, y, z).

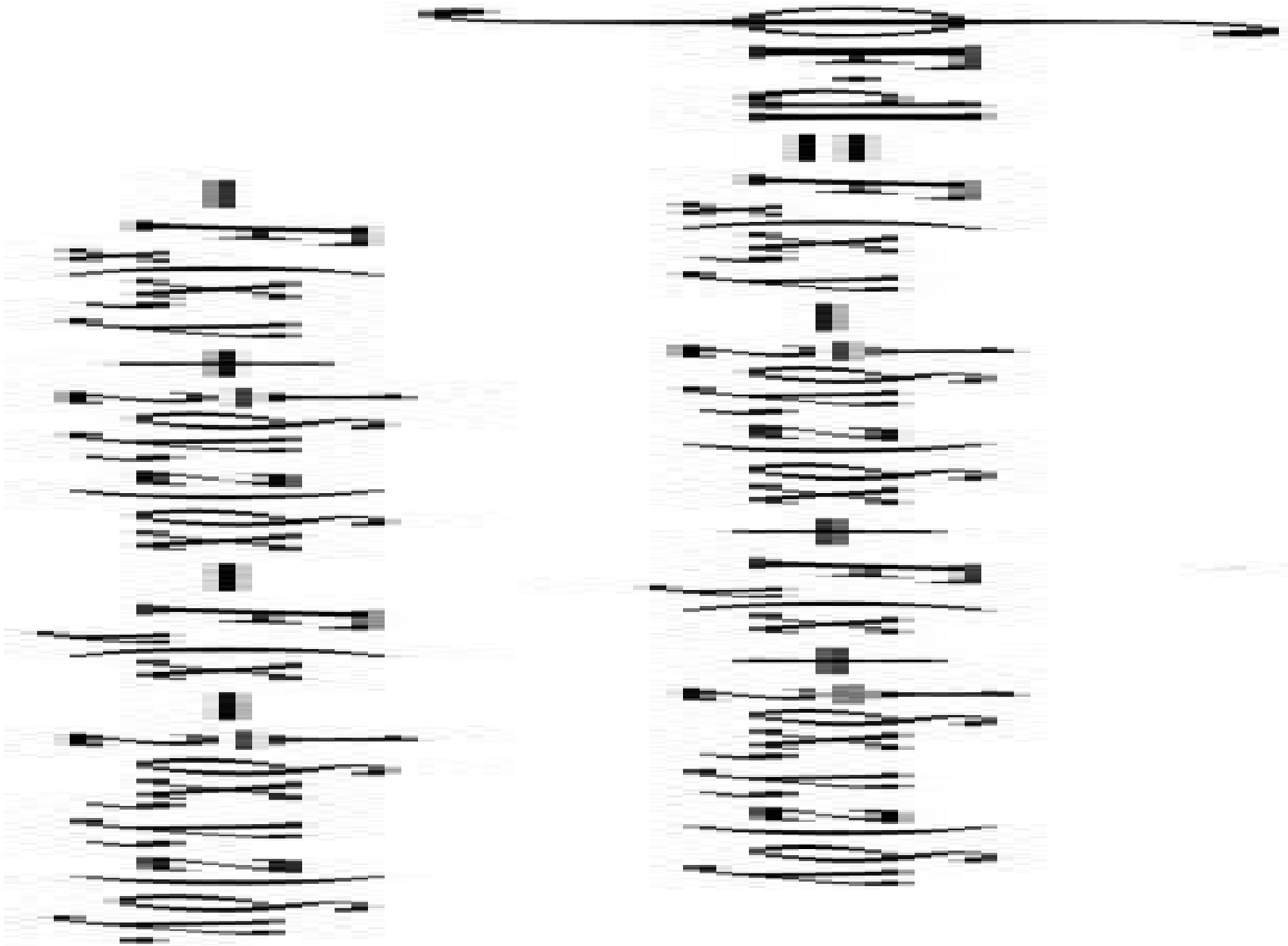

(curl F)z
z
δy
δx
Point (x, y, z)
y
x

rotated in the same sense as the circulation of the path s) is defined by the limit

(4.43)



The operator curl takes different mathematical forms for different systems of coordinates. We will determine its form in a Cartesian coordinate system. The z-component of curl F can be obtained by taking the limit 4.43 over a closed surface in a plane parallel to the x–y plane. Consider the very small surface shown in Figure 4.21, situated at the point with coordinates x, y and z. The sides of the surface have lengths δx and δy parallel to the δx and δy axes respectively. To obtain (curl F)z we have to traverse the path in the direction shown, in which case



Hence

$$(\operatorname{curl} \mathbf{F})_z = \lim_{\delta x \delta y \to 0} \left(\frac{\oint \mathbf{F} \cdot d\mathbf{l}}{\delta x \, \delta y} \right)$$

$$= \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y} \right).$$

Similarly it may be shown that the x and y components of curl F are given in terms of the differentials of the components of F by the equations,

$$(\operatorname{curl} \mathbf{F})_x = \left(\frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z} \right)$$

$$(\operatorname{curl} \mathbf{F})_y = \left(\frac{\partial F_x}{\partial z} - \frac{\partial F_z}{\partial x} \right),$$

We can summarize the three equations for the components of curl F by writing the latter in the form of a determinant.

(4.44)

where ex, ey and ez are unit vectors along the x, y and z axes. The forms that the operator curl takes in some other coordinate systems are given in Appendix B.

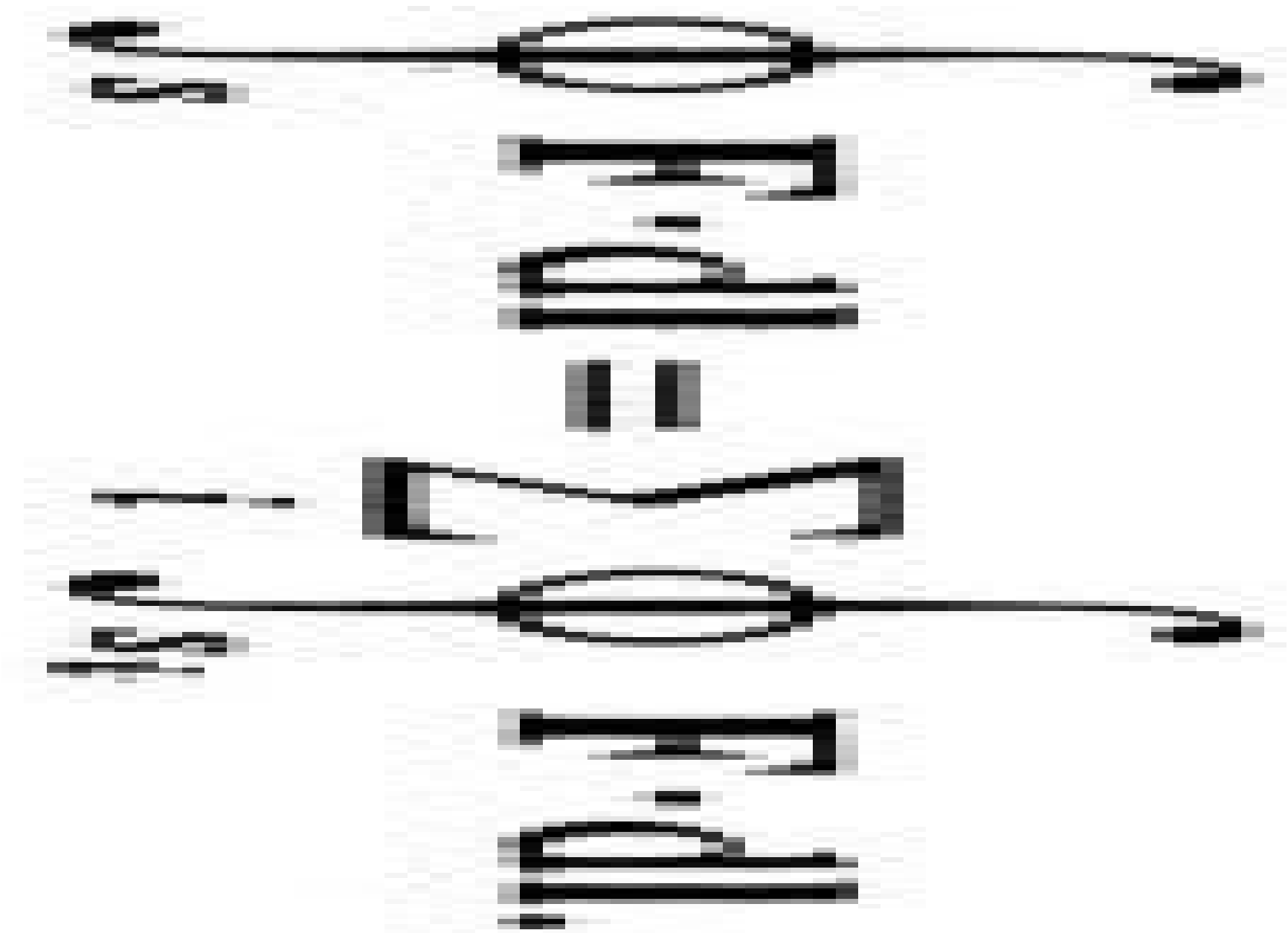
The curl of vector field F is sometimes written in a different notation using the vector operator ∇ introduced earlier in Chapters 1 and 2. There it was pointed out that the divergence of a vector field F, div F, is sometimes written as $\nabla \cdot F$ and that the gradient of a scalar field ψ grad ψ, is sometimes written as $\nabla\psi$ Similarly curl F is sometimes written as $\nabla \times F$. This notation is discussed further in Appendix B. The notation can be misleading when used for the curl of a vector, since at first sight it might seem to imply that the vector $\nabla \times F$ is perpendicular to the vector F, whereas this is not necessarily so.

There is an important theorem called Stokes' theorem which relates the line integral around a closed path of any vector field F to the surface integral of curl F over the surface defined by the path. This theorem states that

(4.45)

and it can be proved by using a construction similar to that used in the current grid. If we split the surface up into elementary surfaces it can be seen that the summation of the line integrals around the perimeters of all the elementary areas

δSi is equivalent to the line integral over the outer perimeter of the whole surface,

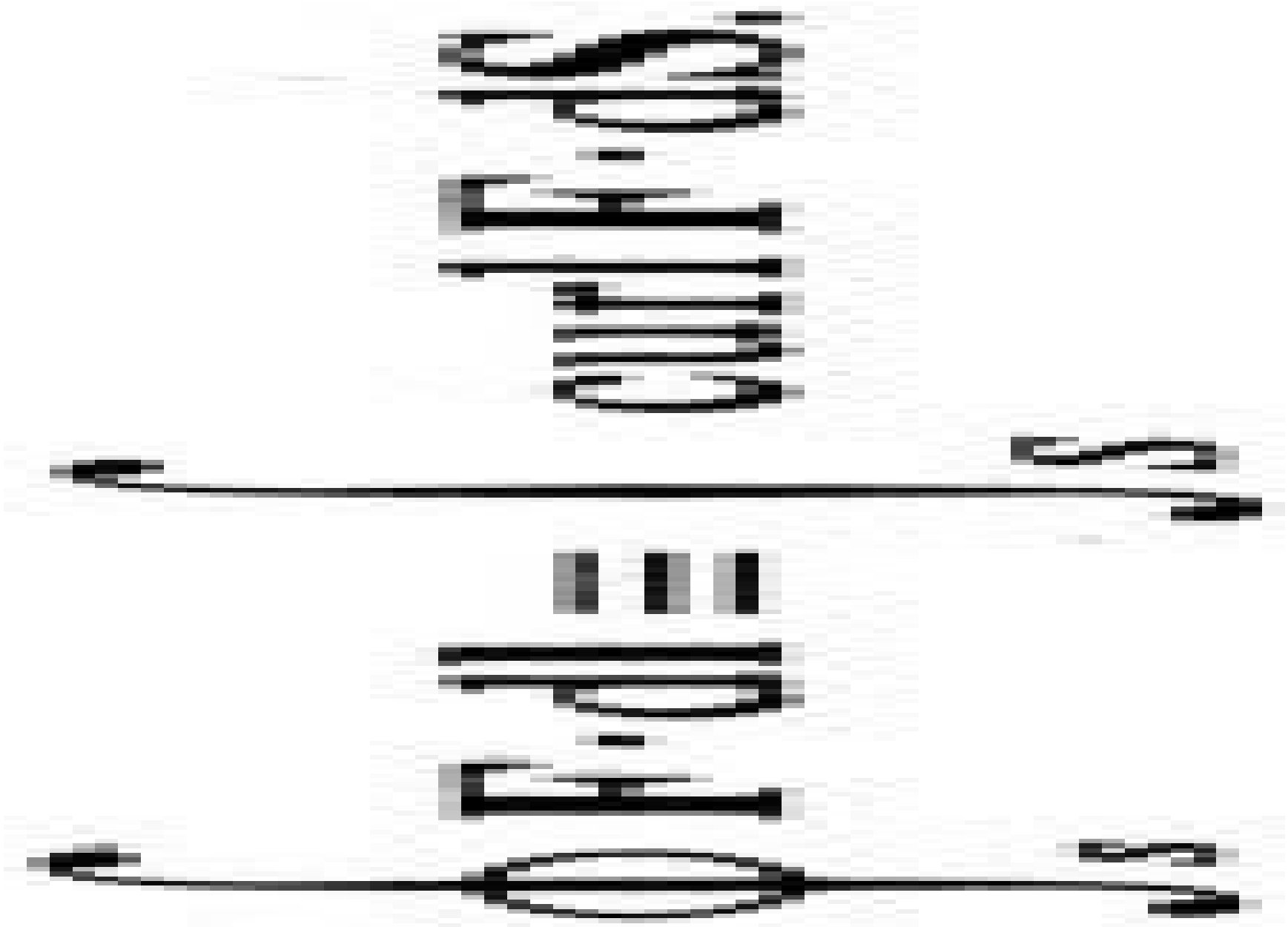


The right-hand side of the equation can be rewritten to give

$$\oint_S \mathbf{F} \cdot d\mathbf{l} = \sum_i \frac{1}{\delta S_i} \oint \mathbf{F} \cdot d\mathbf{l} \delta S_i$$

$$= \sum_i \operatorname{curl} \mathbf{F} \cdot \delta \mathbf{S}_i$$

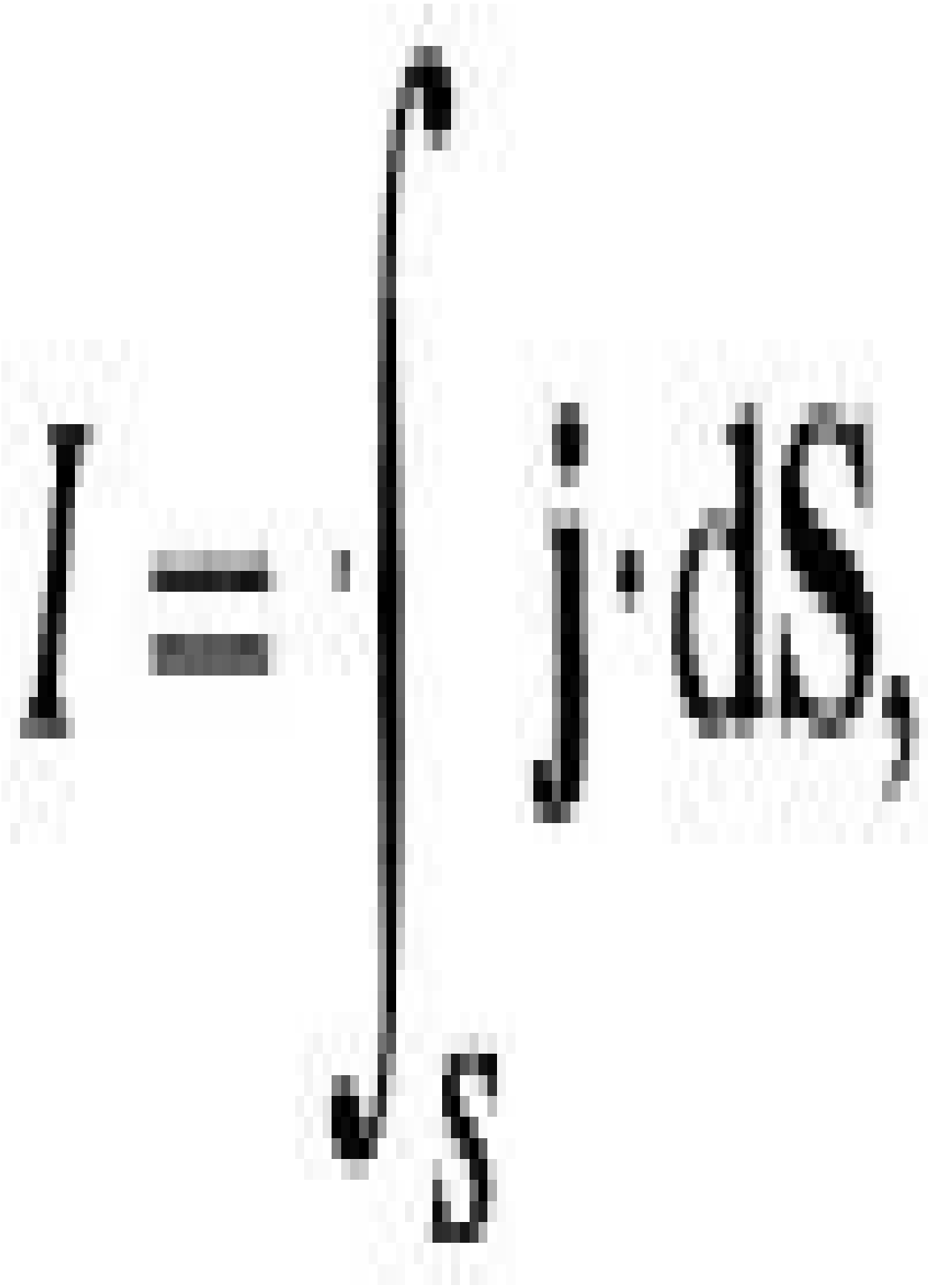
from the definition of curl, Equation (4.43). In the limit of vanishingly small δS_i this leads to Stokes' theorem



It is not easy to visualize the curl of a vector function. If, for example, one is given the form of a magnetic field, only after considerable experience can one quickly answer such questions as ‘Is curl B large, small or zero?’, or ‘In which direction does curl B point?’. Since the components of curl B are all differentials it is obvious that if a vector function is constant everywhere the curl of that function is everywhere zero. Some complicated vector functions however may also have zero curl. For example if we consider the electrostatic field E, since the line integral $\oint E \cdot dl$ over any closed path is zero, it follows from Stokes’ theorem that curl E is always zero.

4.5.2 The vector curl B

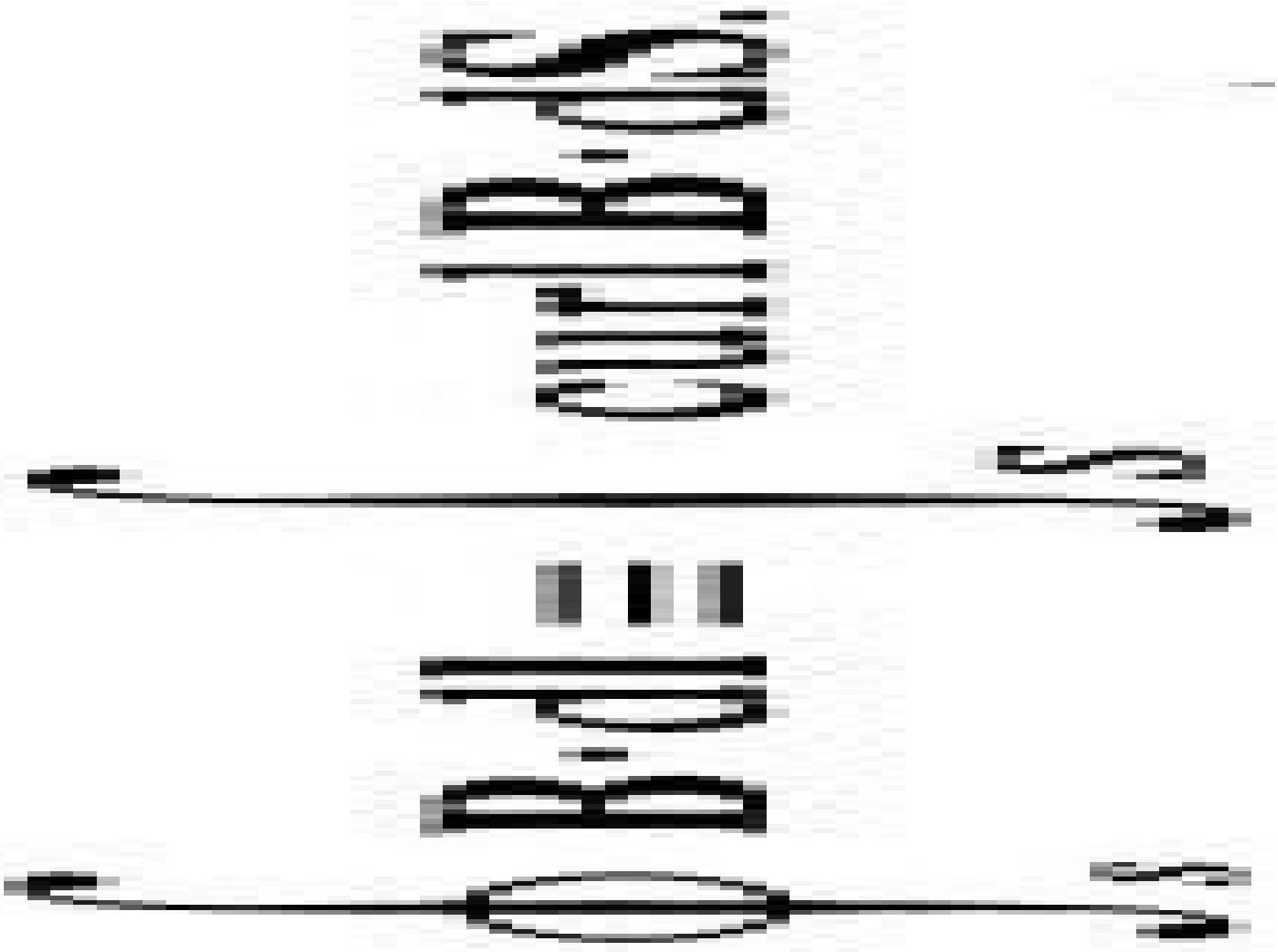
Now that we have introduced the operator curl we can express Ampère’s integral relation 4.31 as a differential equation which the field B must satisfy everywhere. Recalling that the current I through an area S is given by



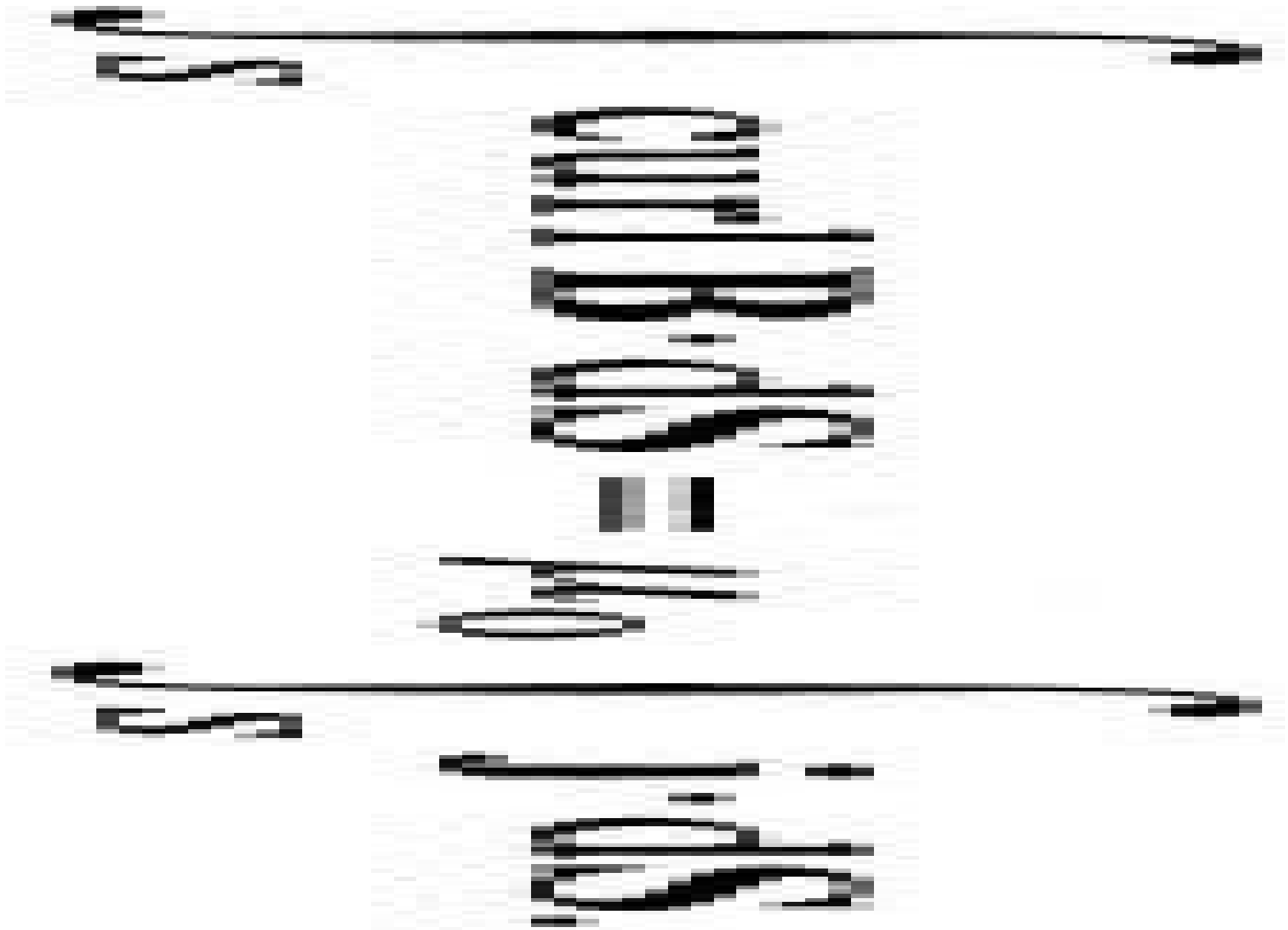
Equation (4.31) can be written

(4.46)

The line integral may be converted into an integral over the surface S by using Stokes' theorem, Equation (4.45),



Substituting in Equation (4.46),



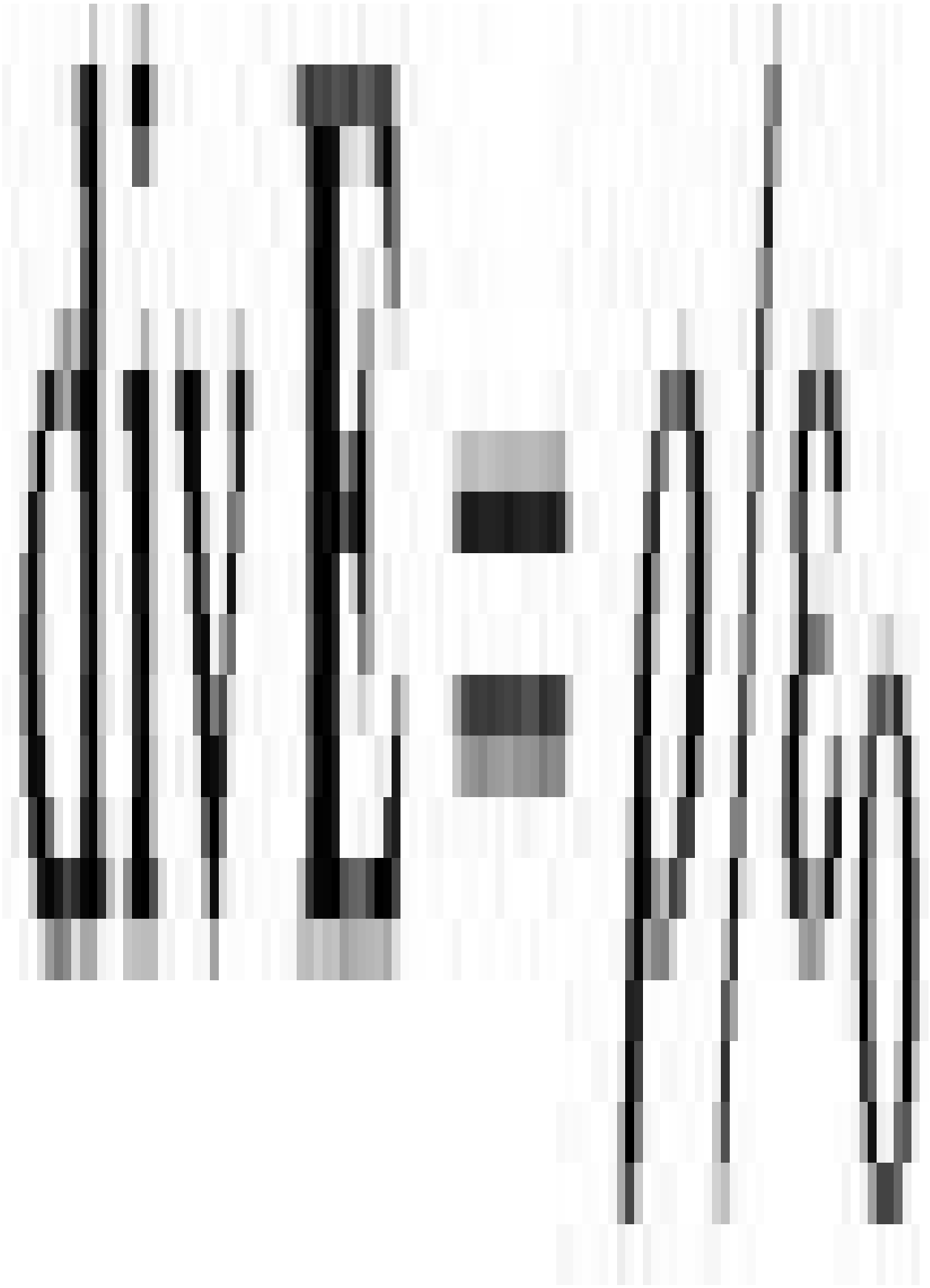
This equation is true whatever the size of the surface S, hence the integrands must be equal everywhere, and

(4.47)

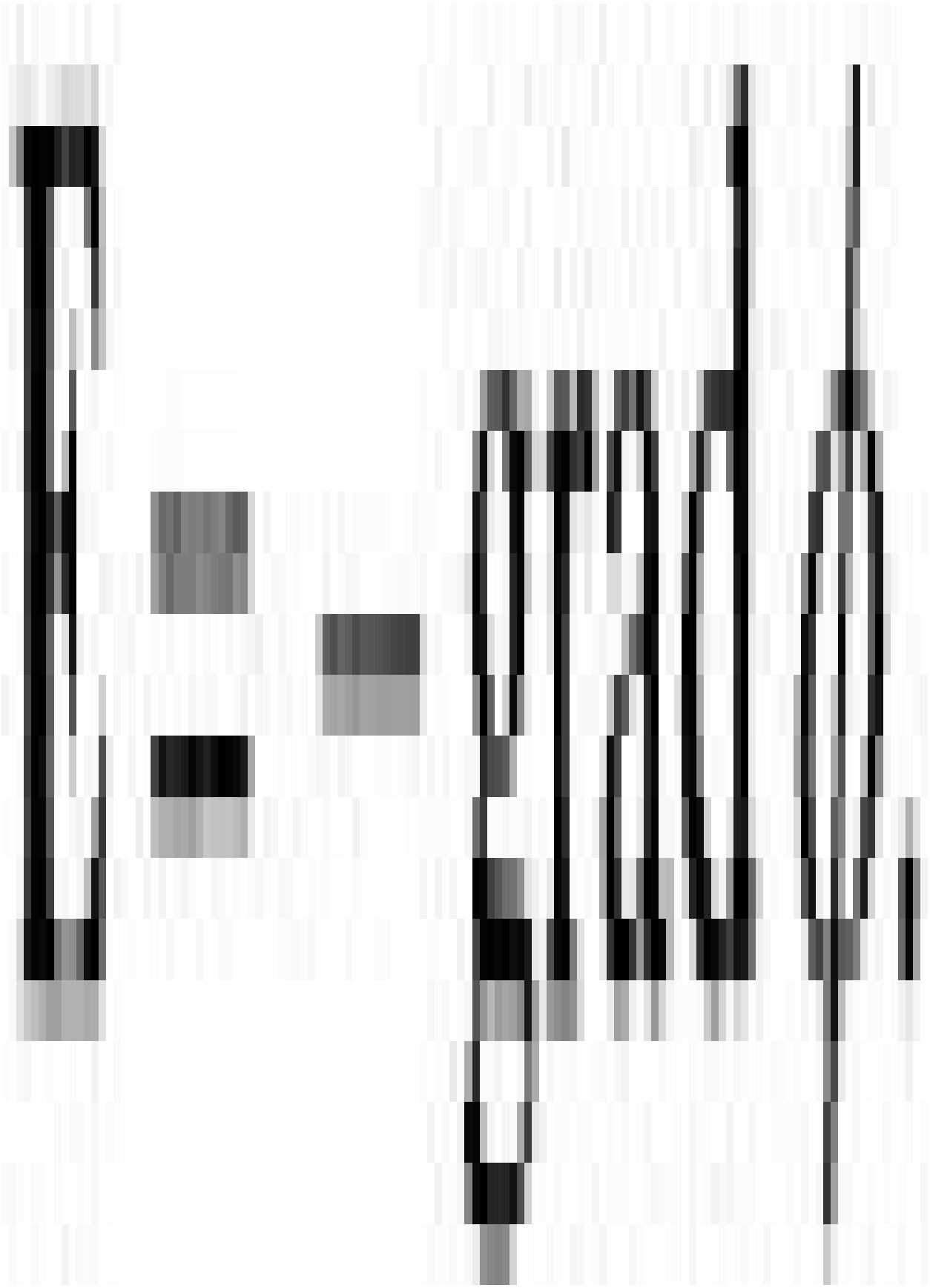
This is Ampère's differential law. It is obeyed at all points by the field B. It relates the field at a point to the current density at that point, and its solution, fitted to the appropriate boundary conditions, will give the magnetic field for a specified current distribution. The general solution to Equation (4.47) for currents in wires is the Biot–Savart law, which can be derived with the aid of the potential introduced in the next section 4.5.3.

4.5.3 The magnetic vector potential

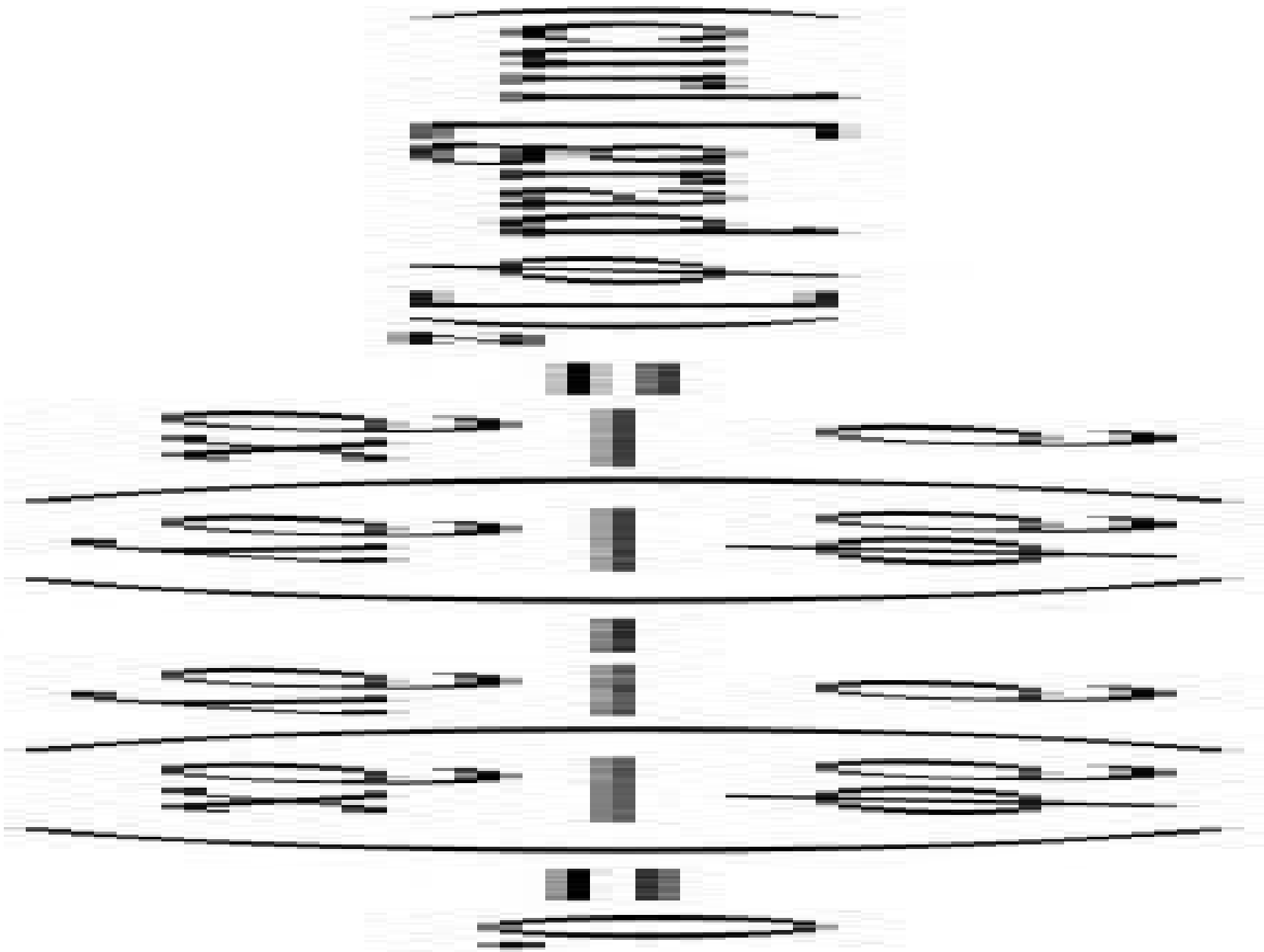
Equation (4.47) plays an analogous role in magnetism to the equation



in electrostatics. The solution of this equation for the field E is made easier by using the electrostatic potential ϕ and the equation



The question arises as to whether there is a potential from which we can always obtain the magnetic field, and whose use would make the solution of Equation (4.47) easier. If a vector function can be obtained everywhere by taking the gradient of a scalar, then the line integral of the vector function around any closed path is zero. It follows that the curl of such a vector function is everywhere zero as we saw in the case of the electrostatic field. This can be proved formally. In Cartesian coordinates the gradient of a scalar function ϕ has components $\partial\phi/\partial x$, $\partial\phi/\partial y$ and $\partial\phi/\partial z$. Hence



Similarly the x and y components of curl grad ϕ are zero, and the scalar satisfies the identity

Thus at points where the current density j is non-zero we cannot derive the field B from a scalar function. We shall now show that the field B can be derived quite generally from a vector potential A, called the magnetic vector potential.

If a vector function of position is obtained from another vector F by taking the curl of F, the divergence of the vector curl F is everywhere zero. Expressed mathematically

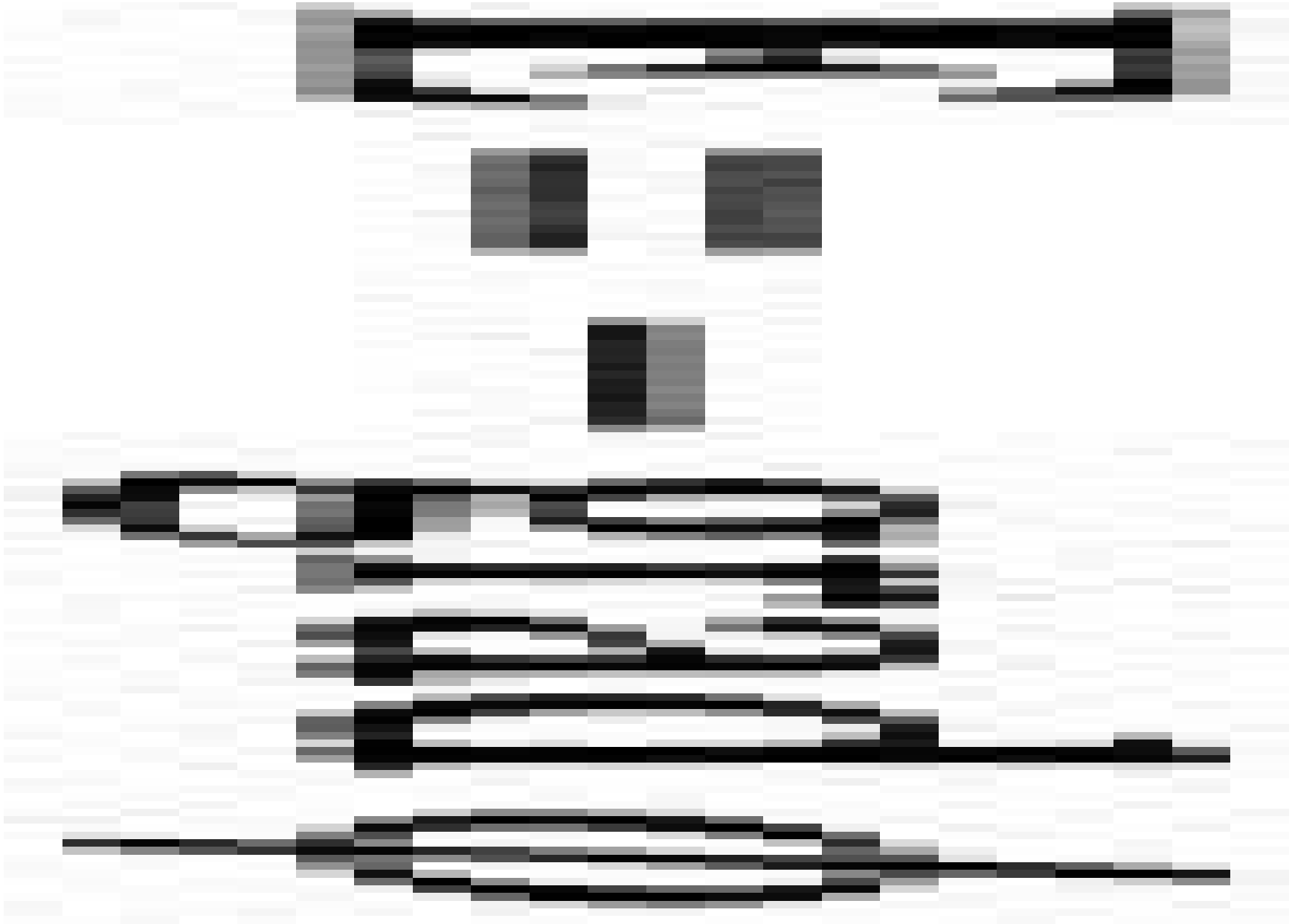
(4.48)

a result which again can easily be proved by writing down div curl F in Cartesian coordinates. (A proof in one particular coordinate system is of course equivalent to a general proof.)

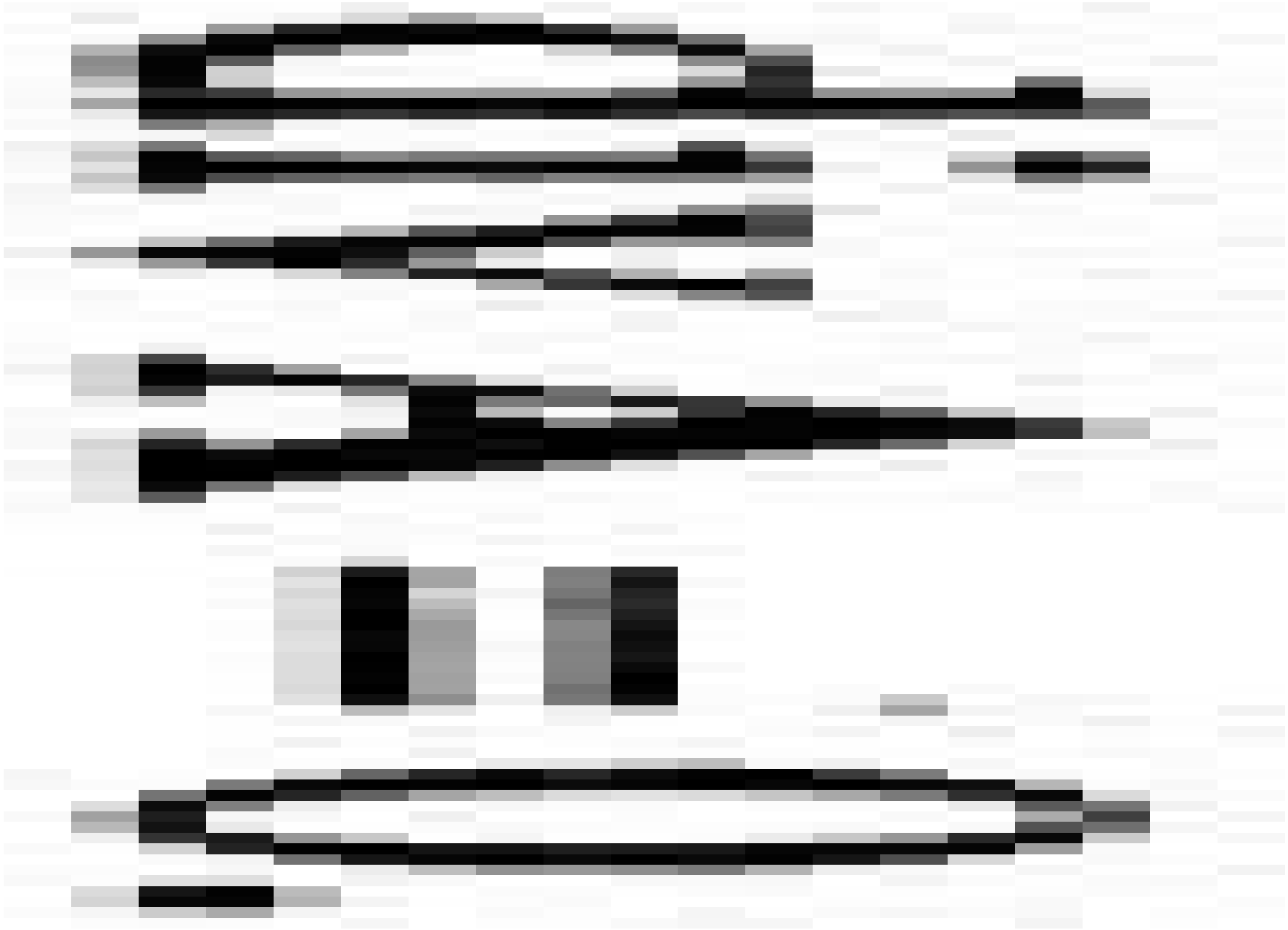
The divergence of the field B is everywhere zero, as we saw in section 4.2.2, and so we can always obtain the field B by taking the curl of another vector field A.

(4.49)

This equation is analogous to the equation



for the electrostatic field. Equation (4.49) does not fix the potential A uniquely. We may add to A any function whose curl is zero and the field B obtained by taking the curl of the resultant addition remains the same. In the same way if we add a constant everywhere to the electrostatic potential ϕ, the field E remains the same. We are free to make additional assumptions about the potentials A and ϕ. The latter is usually defined uniquely by adding the condition that it tends to zero as the distance from a charge distribution tends to infinity. The potential A is usually required to satisfy the condition that, for steady currents,



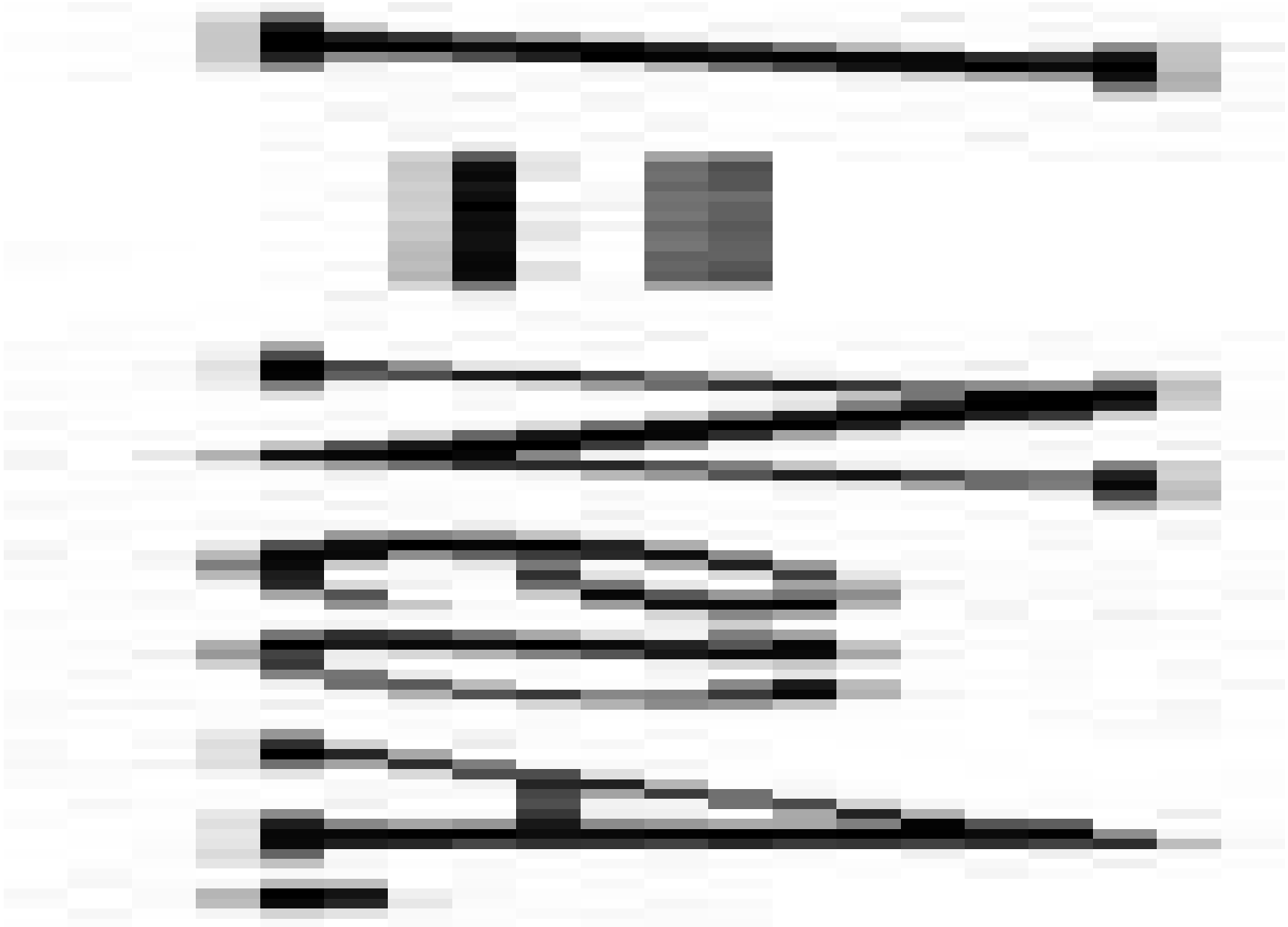
The reason for the choice of this particular condition on the vector potential A is that it usually makes calculations easier than other choices.

Ampère's differential law, Equation (4.47), can be written in terms of the magnetic vector potential A, and the field B for the particular instance of currents in wires can be obtained as the Biot–Savart law. This is done in Appendix C. We do not use the vector A further in this chapter because the Biot–Savart law can be used more conveniently for the steady current problems encountered at the level of this book. However, we will meet it again when discussing fields due to changing currents in Chapter 13.

4.6 FORCES AND TORQUES ON COILS

We now deal with an important practical topip that arises when currents flow in conductors in air. This is the subject of the forces and torques exerted on current carrying coils when they are placed in a magnetic field generated by other sources.

The forces and torques can be calculated in a straightforward manner using the Lorentz force (4.15). We have already considered a rectangular coil in a uniform field in section 4.3.1. More generally, if a steady current I flows in a very small length dl of a conductor whose cross-sectional area is A, then from Equation (4.3),



The number of free electrons in the length dl is NA dl. If the conductor is situated in a magnetic field that has the value B at the position of the element dI, each free electron in that element experiences a force — ev × B. Hence the total force on the element is

$$dF = -N A dl_{ev} \times B$$

$$= I dl \times B_{,}$$

where dl is a vector in the direction of the current in the wire at the position of the element dl. The total force on the whole conductor is obtained by integrating around the closed contour s,

(4.50)

If the position vector of the element dl with respect to an origin of coordinates is r, the total torque on the conductor is

(4.51)

We remind the reader that the field B in the above equations can be taken to be the field from all sources other than the current I in the conductor or can be considered to be the total field. This was discussed in section 4.3.1.

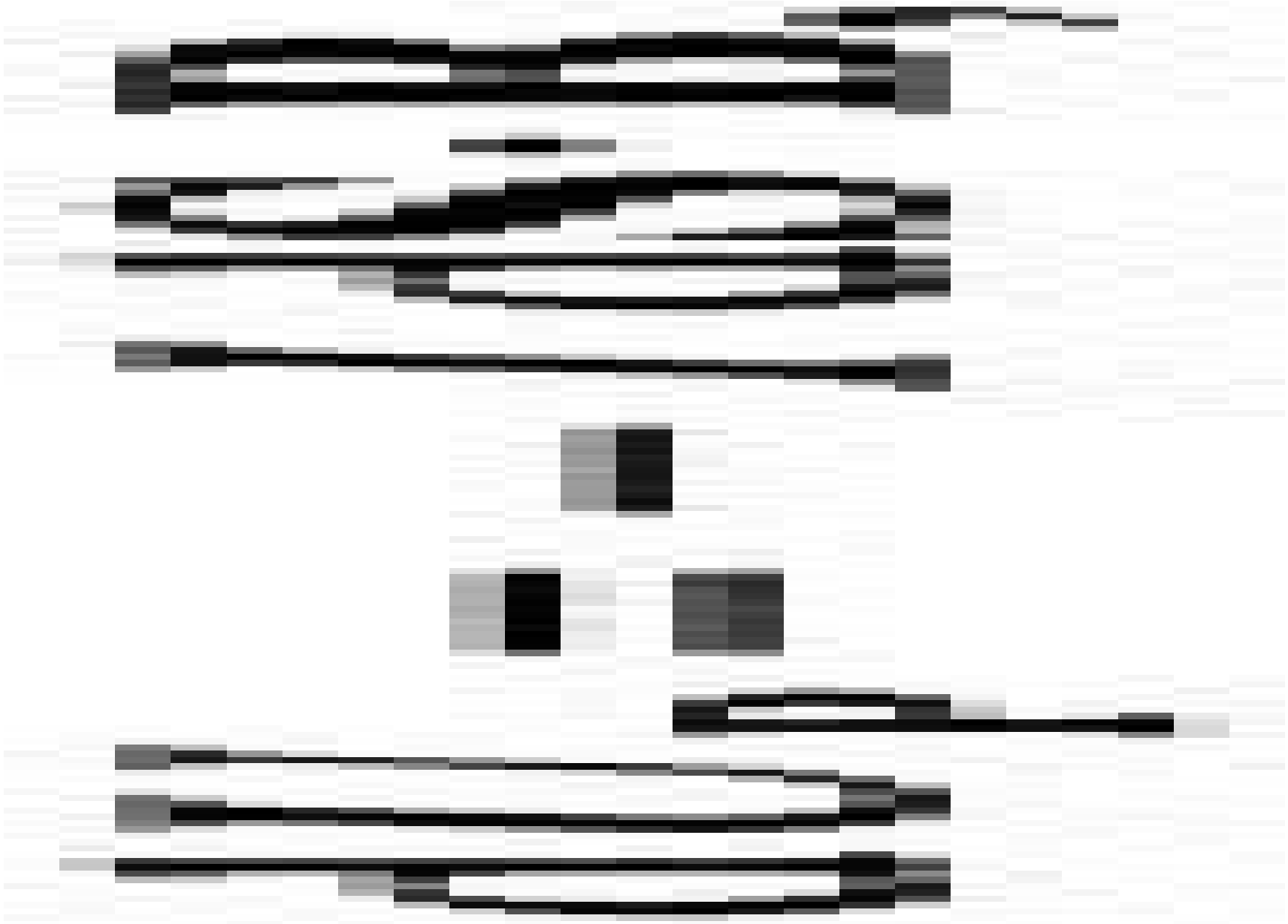
Let us calculate, as a simple example, the force per unit length between two long, thin, parallel wires separated by 4 mm and each carrying a steady current of 50 A. The field of one wire at the position of the other is given by Equation (4.34) to be 2.5×10^{-3} T. The force per unit length is then given by Equation (4.50) to be 0.125 N m^{-1}, i.e. about the weight of 12 grammes at the surface of the Earth. Forces between conductors that carry very large currents can give rise to considerable mechanical problems. There may also be large forces between the separate parts of a single conductor carrying a large current. For example, there cannot be a resultant force on an isolated coil that carries a current, but there will be forces on the individual turns of a multi-looped coil from the current in the other turns. A solenoid wound with gaps between the turns will tend to squash up when a current is passed through it. These effects necessitate considerable care in the design of high current coils.

Expressions 4.50 and 4.51 can be used quite generally to calculate forces and torques on current carrying coils, although the calculations are often tedious.

4.6.1 Magnetic flux

In this section we will derive alternative expressions for the forces and torques on coils in terms of the magnetic flux through them. These expressions are very useful and often easier to use than Equations (4.50) or (4.51). We will first derive them using the idea of the current grid and its equivalent sheet of magnetic dipoles, and leave the definition of magnetic flux until afterwards.

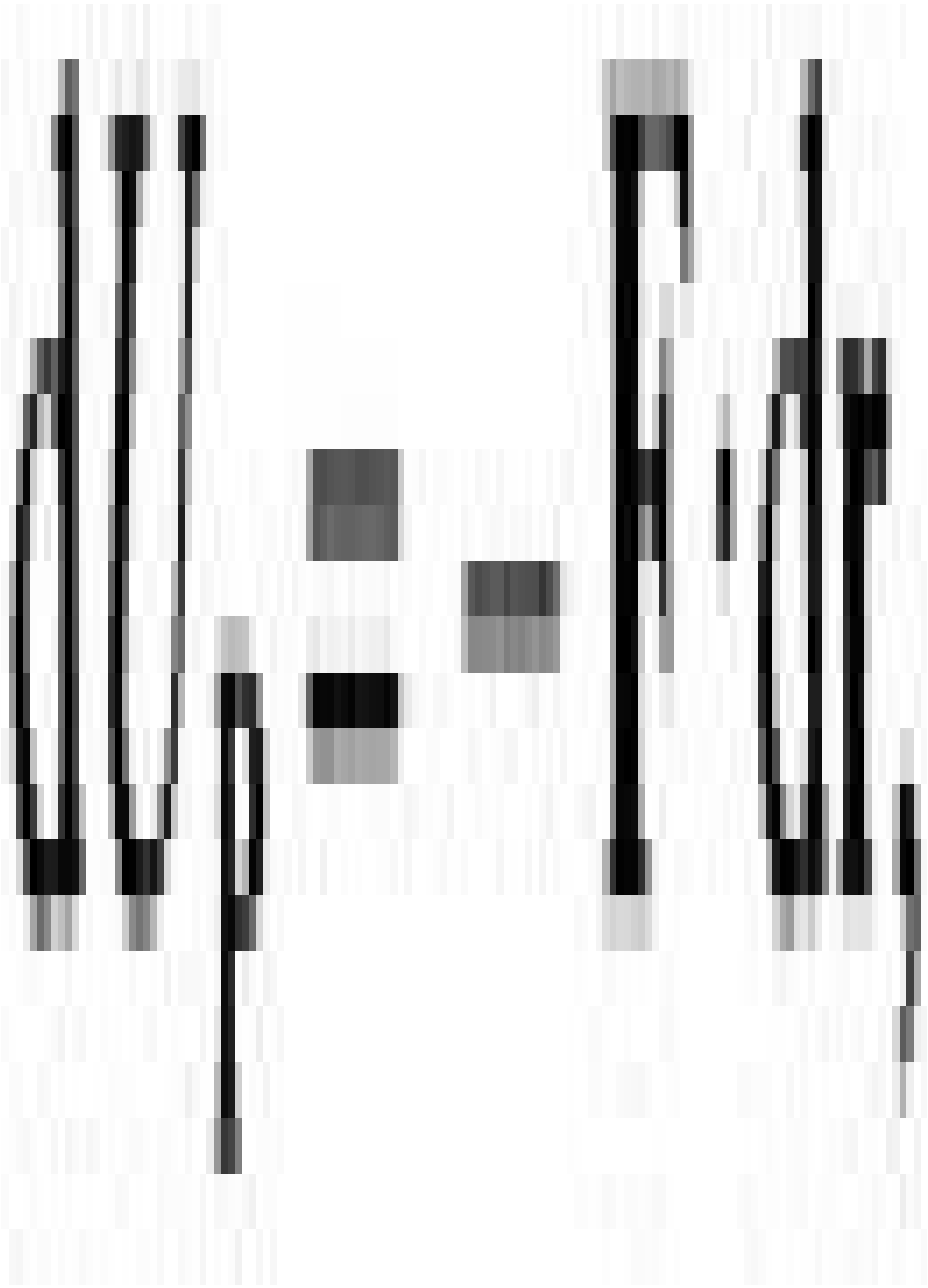
Consider a closed coil which carries a steady current I. We have seen in section 4.4.1 that the magnetic effects of the current are the same as those of a current grid over the surface S of the coil, and each elemental area dS of the surface behaves as though it were a magnetic dipole of moment I dS. If B is the magnetic field at the position of an element dS that arises from sources other than the current I, the dipole I dS has a potential energy



from Equation (4.23). The potential energy* of the whole coil in the external field is

(4.52)

If the coil with the current I constant is displaced a distance dr while a magnetic force F acts on it, the change of potential energy is given by



and

$$F = -\operatorname{grad} U_p$$

$$= I \operatorname{grad} \int_S dS \cdot B,$$

from Equation (4.52). This equation is usually written as

(4.53)

Where

(4.54)

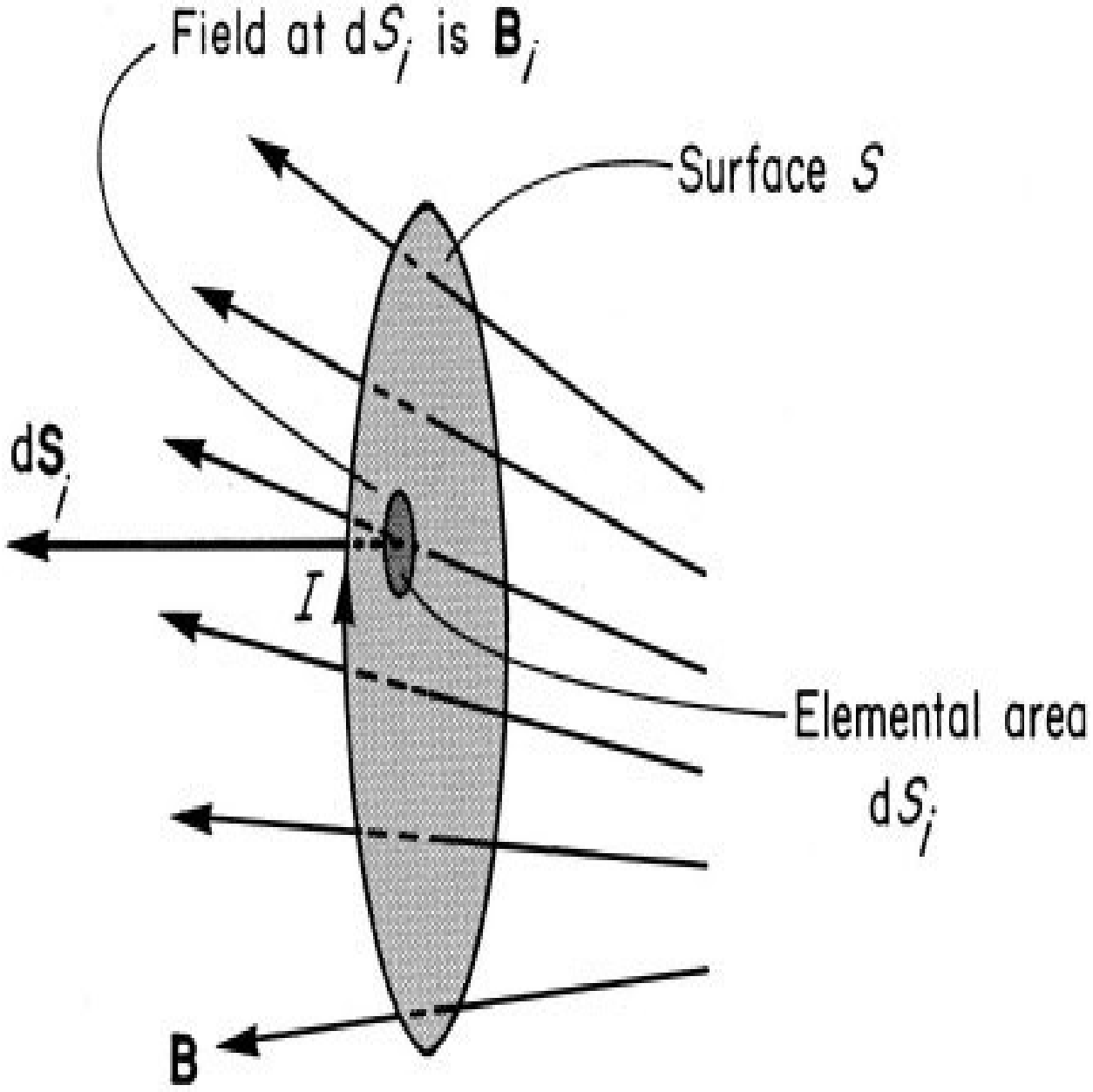
defines the magnetic flux through the circuit. Since the lines of the field B are continuous the surface integral of the field is the same over any surface whose edge is defined by the coil, i.e. the magnetic flux through any such surface is the same. The sign of the magnetic flux Φ in Equation (4.53) is determined by the directions of the current in the coil and the magnetic field. The direction of the vectors dS in Equation (4.54) is determined by the right-hand screw rule used with the current in the coil giving the screw direction, as illustrated in Figure 4.22. The sign of the magnetic flux is then positive or negative depending on the direction of the field B. The dimensions of magnetic flux are [mass] × [charge] -1 × [time]-1 × [length]^2. The SI unit of magnetic flux is called the weber (Wb). The unit of magnetic field B is sometimes called the weber per square metre (Wbm-2), i.e. 1 weber per square metre ≡ 1 tesla.

The torque on a coil situated in a magnetic field and carrying a current I is given by a formula analogous to Equation (4.53). For example if the coil can rotate about a vertical axis in its plane, and the angle between the plane of the coil and a fixed horizontal line is θ, the size of the torque T on the coil about the vertical axis is given by

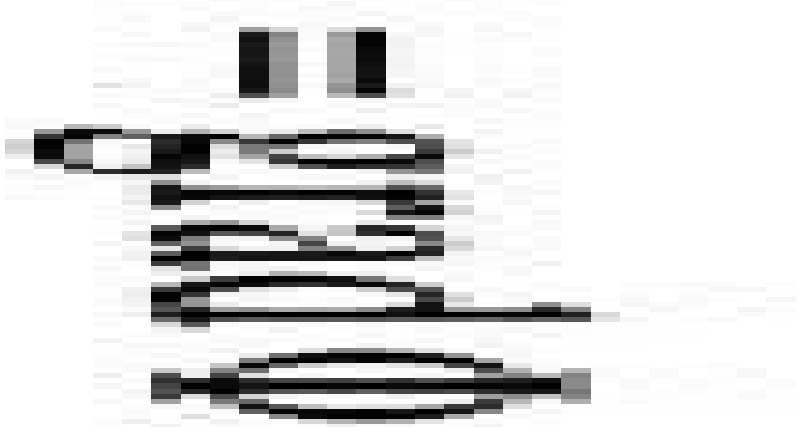
(4.55)

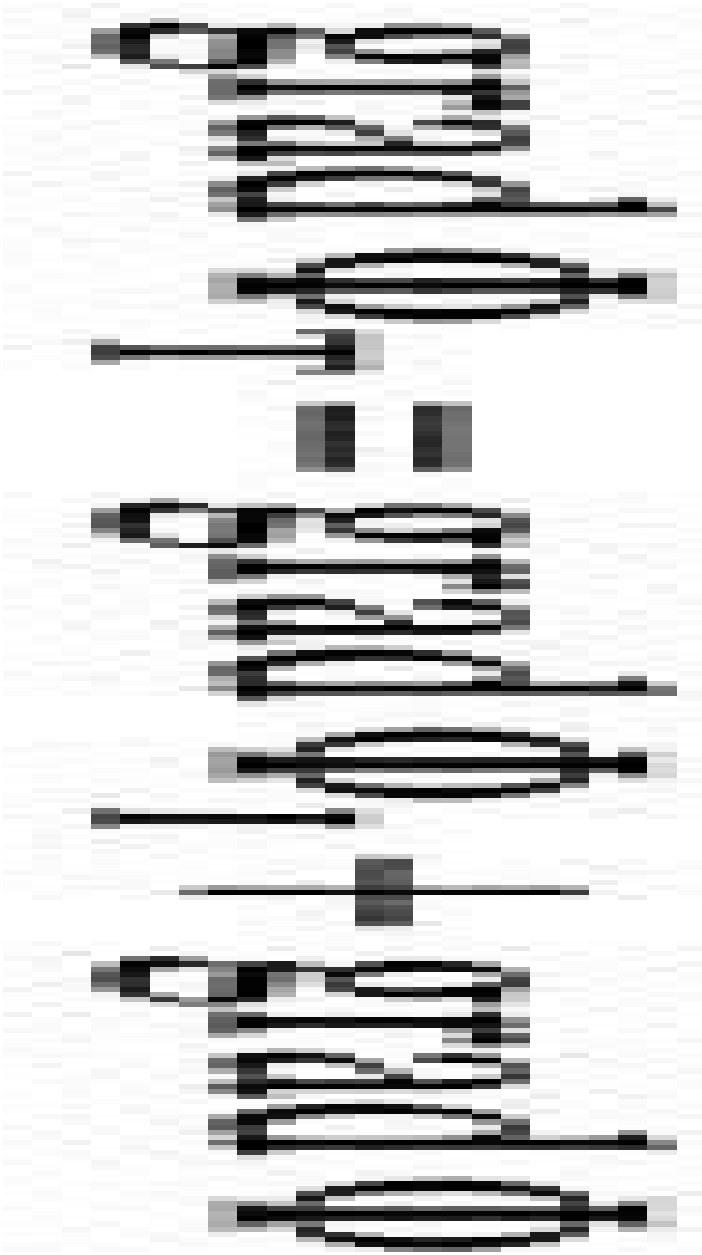
Equations (4.53) and (4.55) again make it clear that the force and torque on a current-carrying coil can be expressed equally well in terms of the total magnetic field or in terms of the magnetic field due to sources other than the current in the coil itself. If the magnetic flux Φ used in Equations (4.52) and (4.54) is calculated from the total magnetic field, the resulting values of the force and torque are unchanged. The total flux ΦT can be written as the sum of the flux Φ1 due to the current in the coil, and the flux Φ due to the field B arising from sources other than the current I. Now

Figure 4.22. The magnetic flux through a surface in a magnetic field.



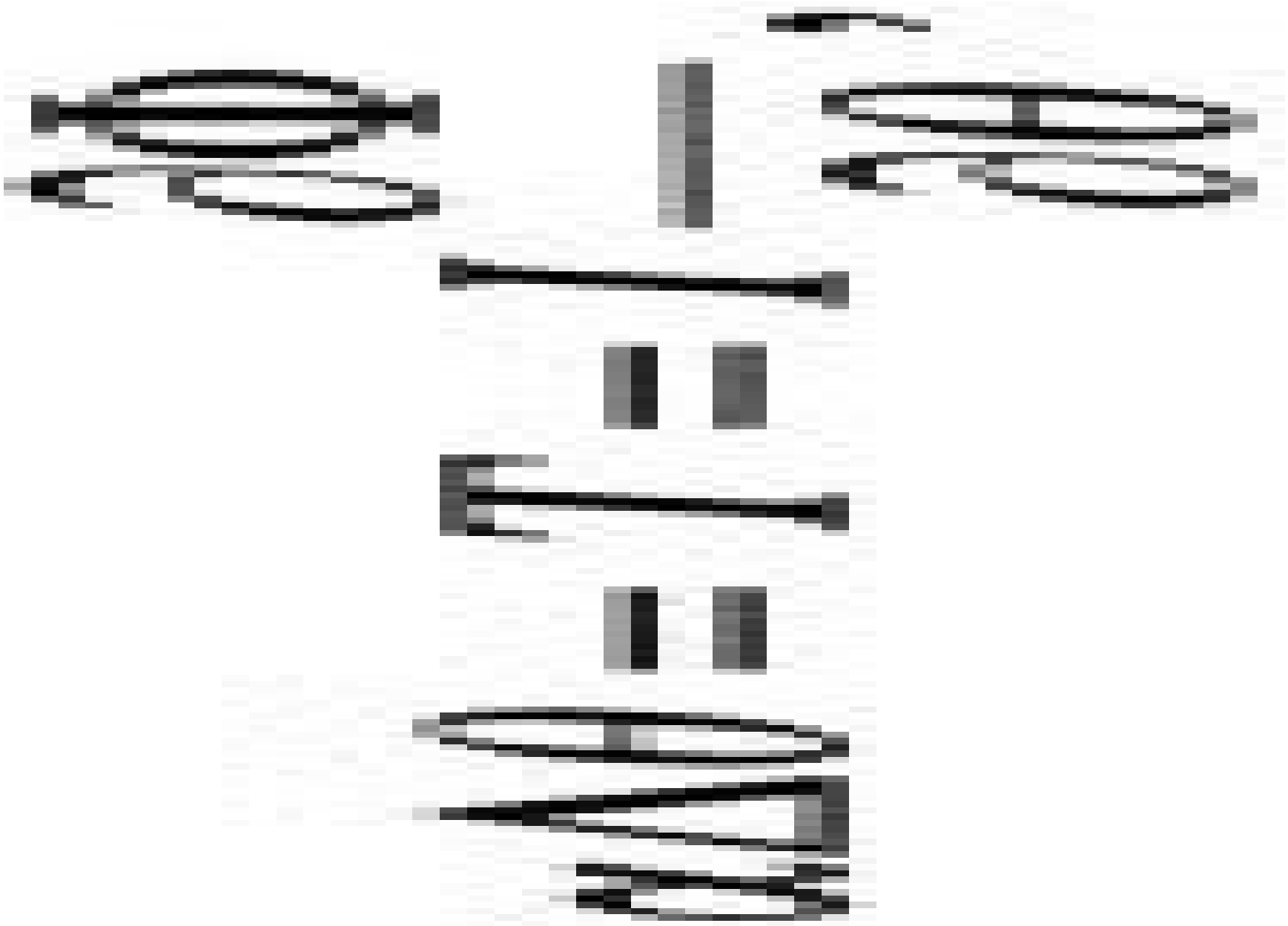
$$\text{Flux through } S = \sum_i \mathbf{B}_i \cdot d\mathbf{S}_i = \int_S \mathbf{B} \cdot d\mathbf{S}$$



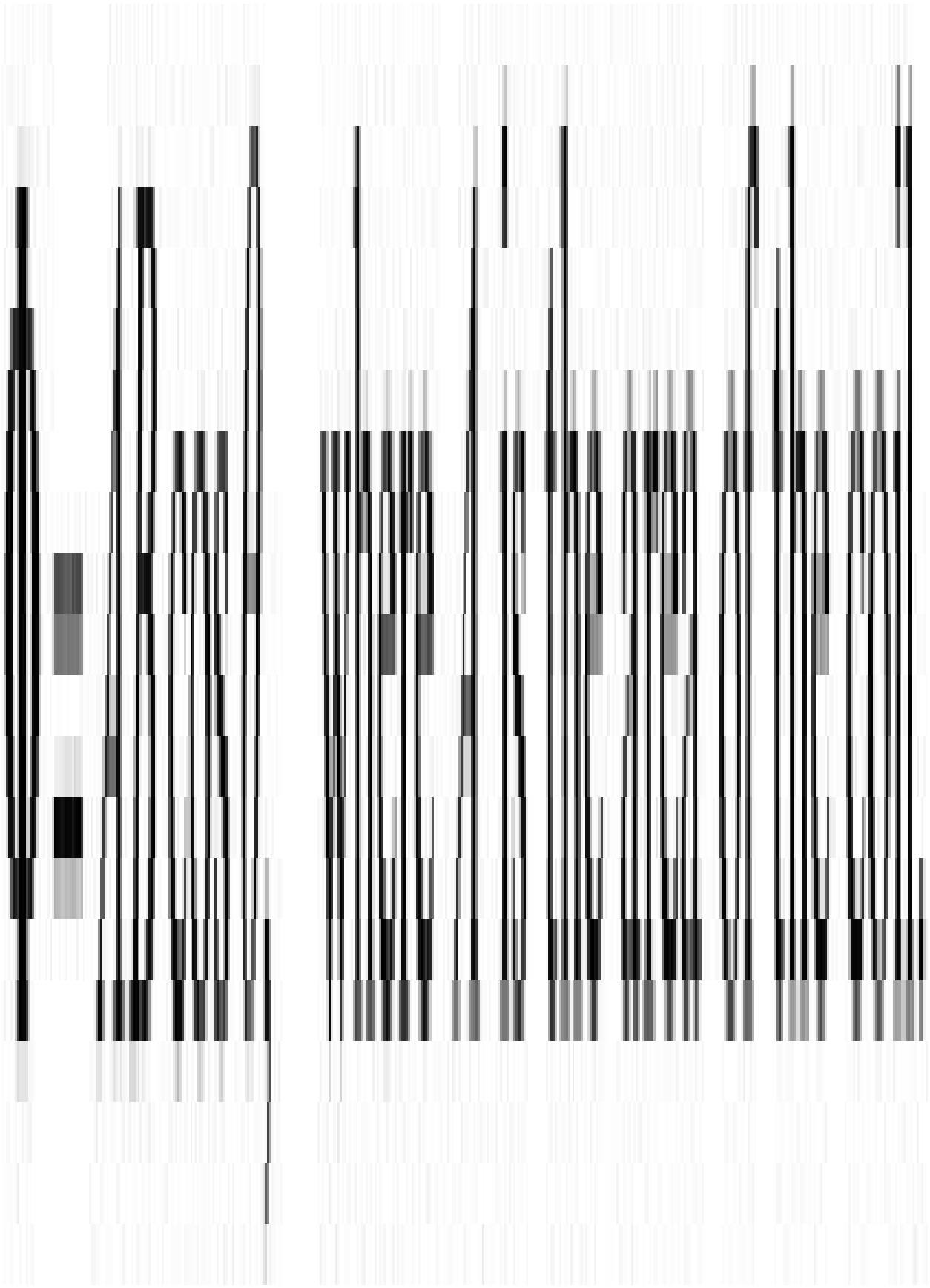


since the flux $\Phi 1$ will not change when the coil is displaced slightly at constant current I. Hence the calculated force on the coil (and similarly the torque) remains the same whether ΦT or Φ is used in the equations.

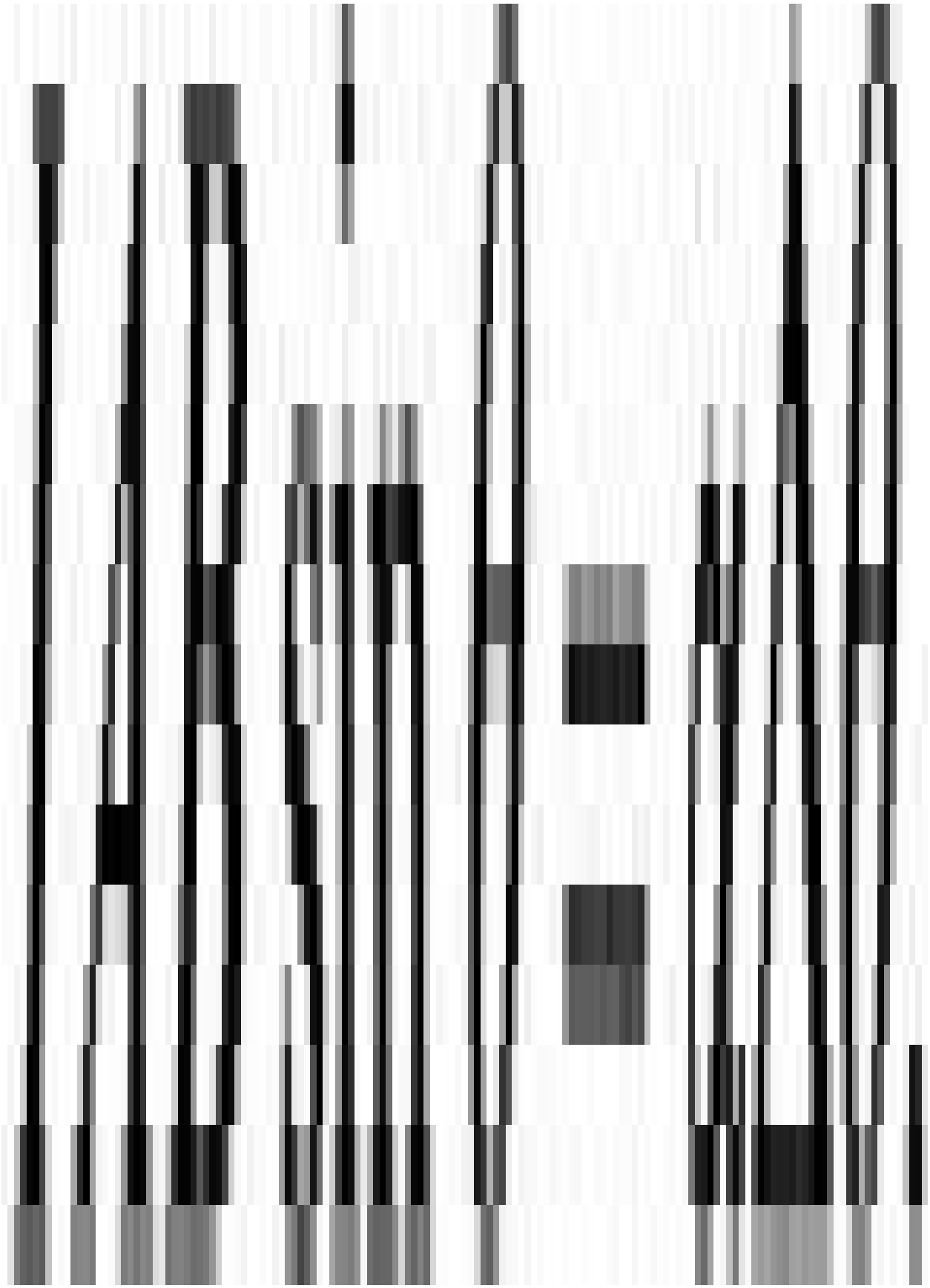
An example illustrating the above ideas is given by a coil suspended in a magnetic field such as the coil shown in Figure 4.9. With no current in the coil the suspension holds it steady at some equilibrium position given by an angle $\theta 0$ which measures the rotation around the axis of the suspension. When a current I is passed through the coil it experiences a torque and is deflected to a new angle θ at which the torque T on the coil due to the magnetic forces balances that in the opposite direction from the suspension. The size of the deflection, $\Delta\theta$, is thus given by



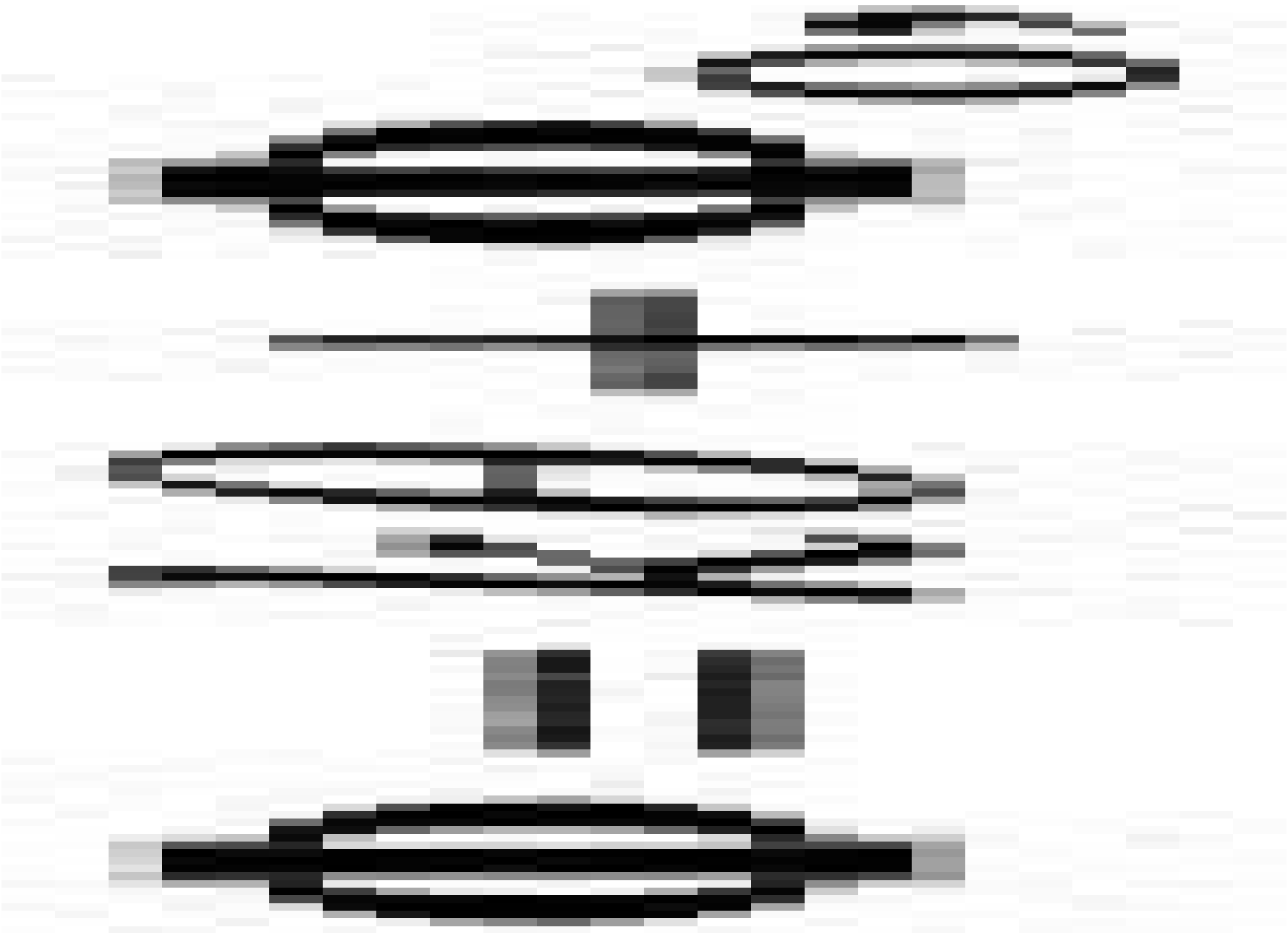
where α is the restoring torque per unit twist provided by the suspension. Referring to Figure 4.9 we see that



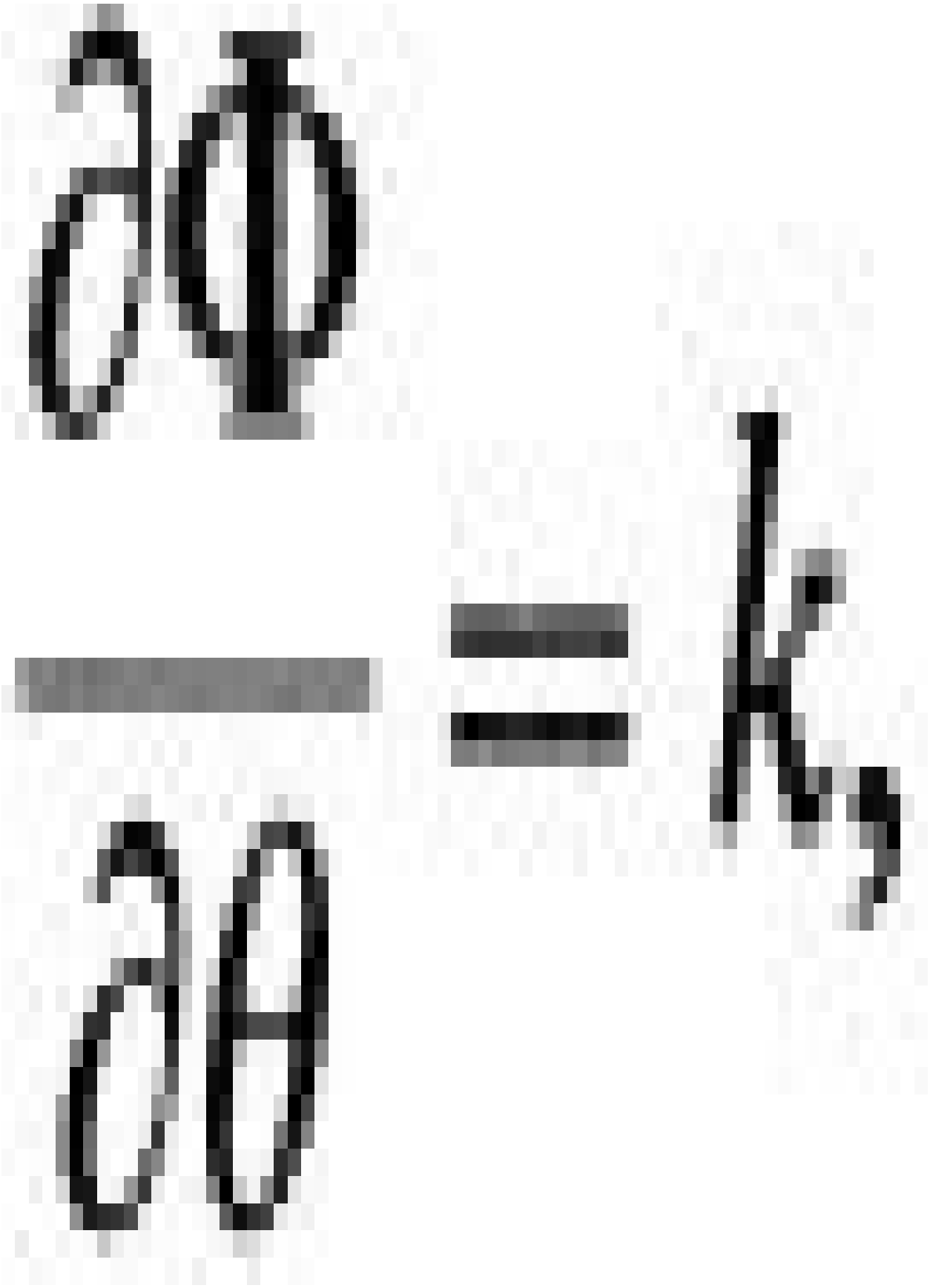
The magnitude of $\partial\Phi/\partial\theta$ is thus equal to AB sin θ, and hence



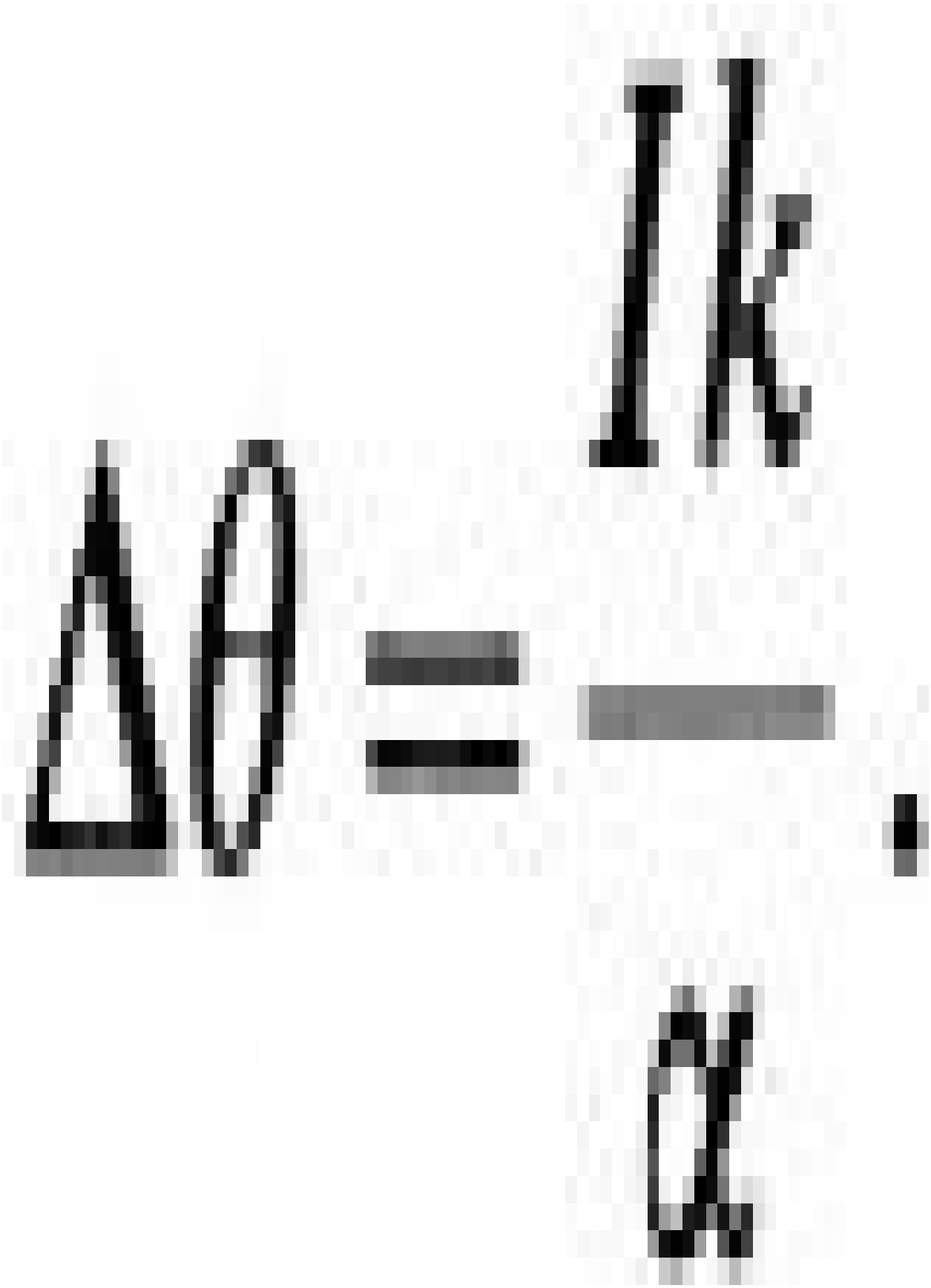
If the field in which the coil sits is shaped so that Φ varies linearly with θ over a small range of deflections $\Delta\theta$, then



where k and Φ0 are constants,



and



This is the principle on which moving coil instruments are based. These instruments were used for measuring small currents, and can still be encountered in teaching laboratories. Nowadays currents are usually measured in the laboratory with electronic instruments which measure and amplify the voltages produced by the currents as they pass through resistors.

4.7 THE MOTION OF CHARGED PARTICLES IN ELECTRIC AND MAGNETIC FIELDS

The force on a particle of charge q moving with velocity v in electric and magnetic fieldsE and B is given by Equation (4.16):

If the particle is moving in a vacuum at the surface of the Earth there are no nearby molecules with which it can collide or interact, and the only force acting on the particle, other than that due to the electric and magnetic fields, is the force of gravity. When one is considering atomic particles it can easily be shown that the gravitational forces are usually negligible. For example a proton moving with a speed of $2 \times 10^5 \text{cms}^{-1}$ (which is roughly the speed of hydrogen molecules at room temperature) perpendicular to the Earth's magnetic field experiences a force about 10^6 times greater than the gravitational force. The electric force on a proton in a field of 10^{-7} V m^{-1} is about equal to the gravitational force. If we neglect the gravitational force, Newton's law for the particle in vacuum becomes

(4.56)

where p is the momentum of the particle. For given fields E and B one must solve Equation (4.56) to find the position of the particle as a function of time.

4.7.1 The motion of a charged particle in a uniform magnetic field

A simple example which is often encountered is that of the motion of a charged particle of mass m in a uniform magnetic field. Since the magnetic force always acts at right angles to the motion, the force cannot increase the energy of the particle, and its speed will remain constant. If the particle is moving in a plane perpendicular to the direction of the uniform magnetic field B, the force $qv \times B$ on the particle is constant in magnitude and always at right angles to the direction of motion. The particle therefore moves in a circle of radius R, where

(4.57)

or

(4.58)

This is illustrated in Figure 4.23.

It can be seen from Equation (4.58) that particles of the same ratio of momentum to charge move in circles of the same radius in a uniform magnetic field. Particles of a given value of (mv/q) are often selected by accepting only those that traverse a path of fixed radius when moving in a uniform field. Figure 4.24 shows abeam of charged particles, all with the same mass and charge but having a spread of velocities. We consider a somewhat unrealistic beam in which all the particles are moving in the same direction without any component of velocity perpendicular to this direction. The beam enters through

Figure 4.23. The path of a positively charged particle moving perpendicular to a uniform magnetic field. The magnetic force is perpendicular to both the field and the velocity of the particle, and the path is a circle.

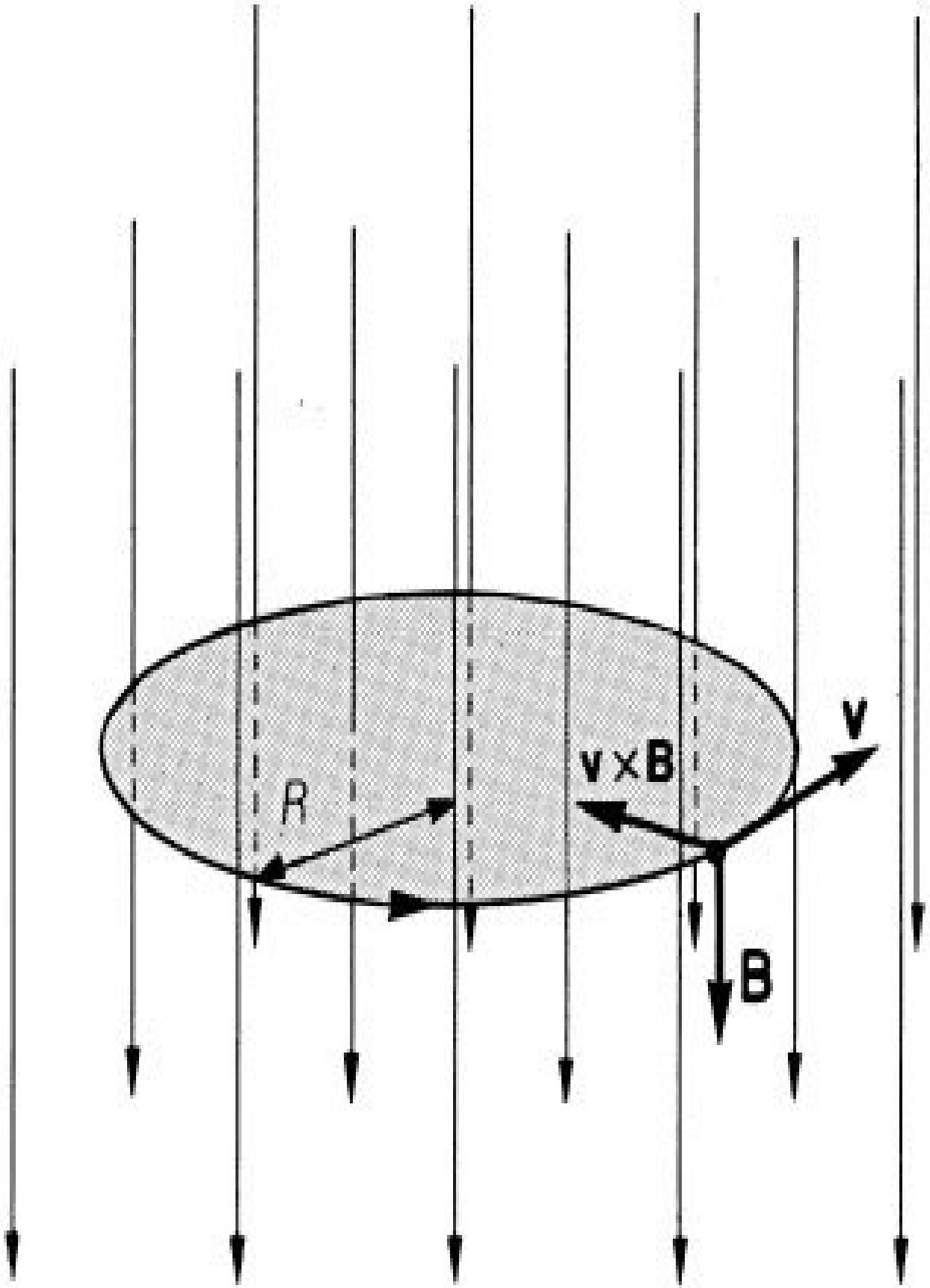
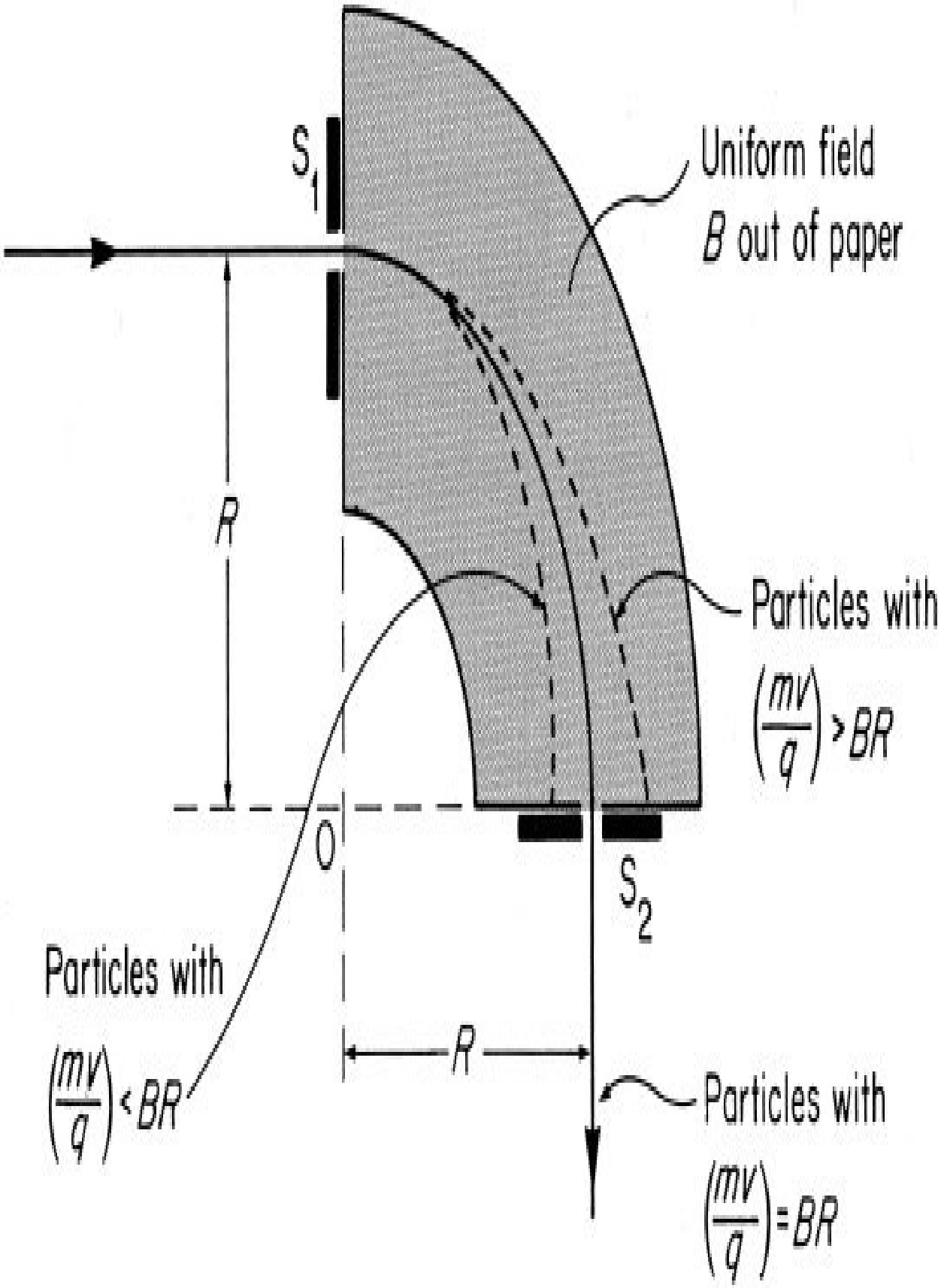
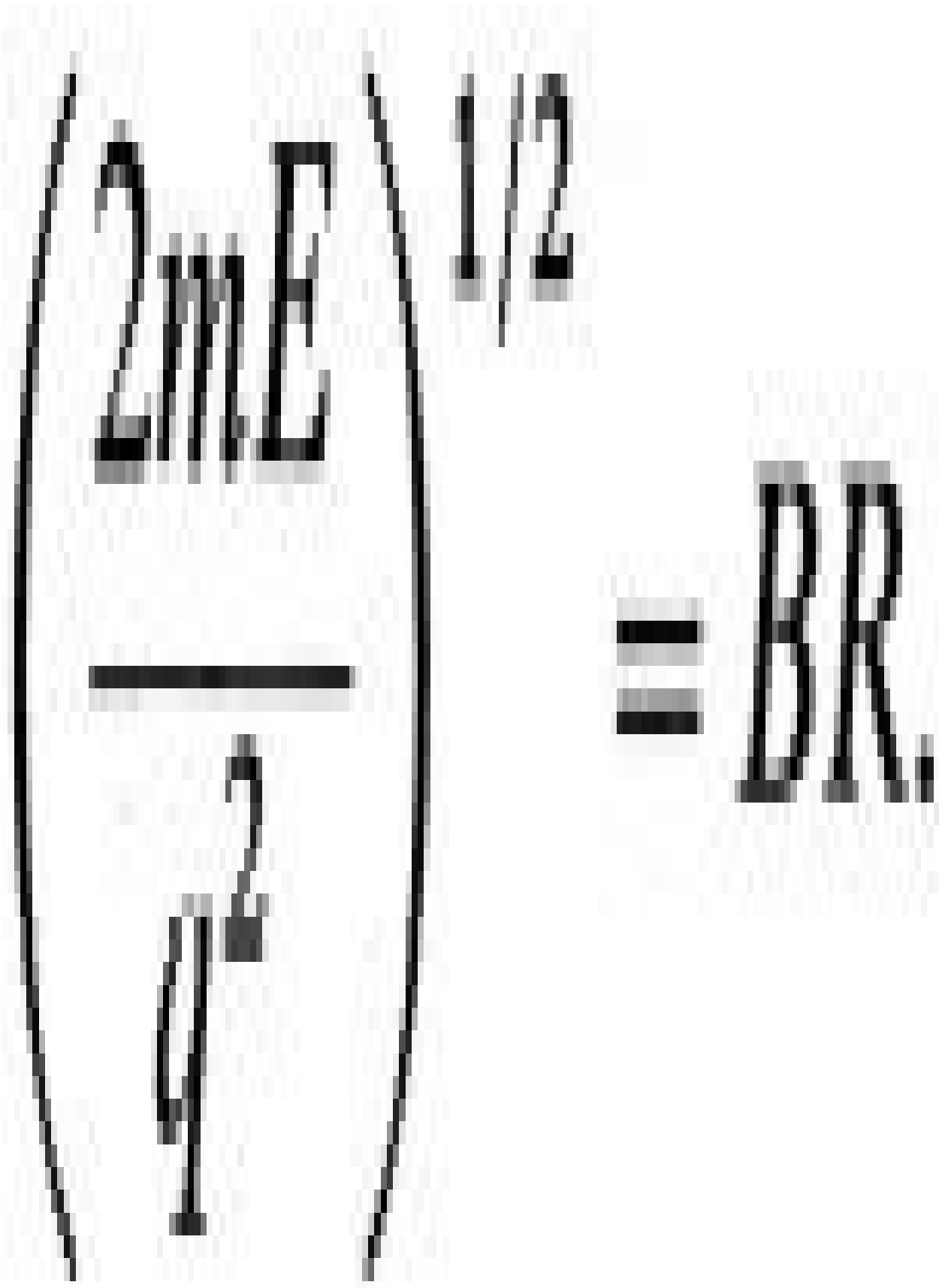

v
v×B
R
B

Figure 4.24. The trajectory of a positively charged particle entering a uniform magnetic field at slit S1 and leaving at slit S2. There is a uniform magnetic field out of the paper over the shaded region. Particles with different speeds are bent in circles of different radii.

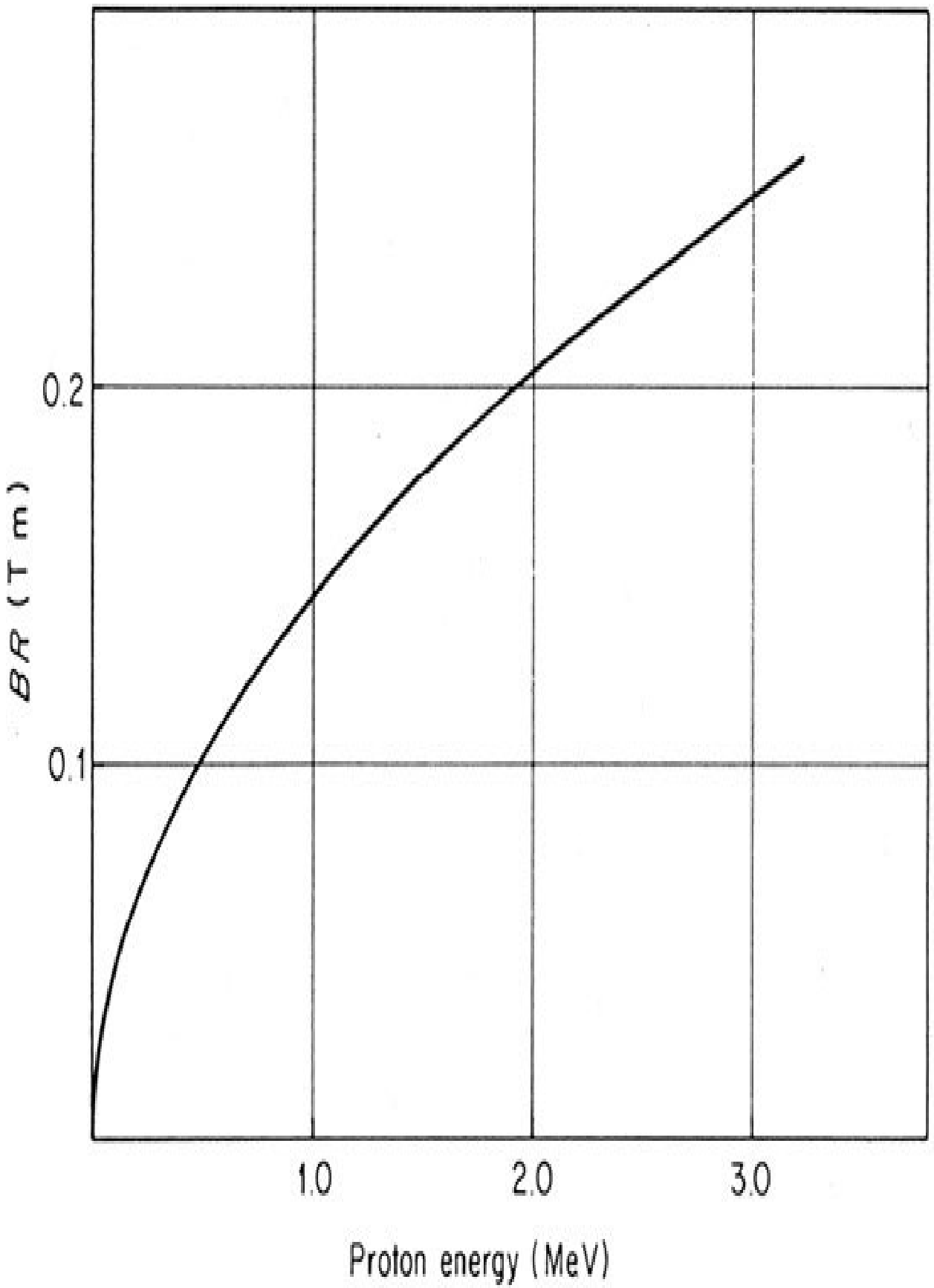

S_1
Uniform field
B out of paper
R
Particles with
$\left(\frac{mv}{q}\right) > BR$
O
S_2
Particles with
$\left(\frac{mv}{q}\right) < BR$
R
Particles with
$\left(\frac{mv}{q}\right) = BR$

a narrow slit S1 into a region where there is a uniform magnetic field in a direction perpendicular to the velocity of the particles. The defining slit S2 allows particles to leave the field only if they have moved in an arc of a circle centre O, where OS1 = OS2 = R. The selected particles then have the value of (mv/q) given by Equation (4.58)*. An equivalent statement is that the slits and magnetic field together select those particles which have a constant value for the quantity (mE/q^2), where E is the kinetic energy of the particle. This technique is used in one type of mass spectrometer. Mass spectrometers measure, for example, the chemical composition of a substance by mass analysis of singly charged ions formed from its constituents. The ions, with charge q = e, are accelerated over a potential difference V volts to a kinetic energy V electron volts. They are then passed through a uniform magnetic field such as provided by the magnet illustrated above. This type of magnet is called a dipole magnet. Varying the magnetic field B gives a yield curve of ions of different mass. The same technique can be used to measure the energies of particles from accelerators. From Equation (4.58)



The quantity BR is called the magnetic rigidity of the particles and for protons is plotted against energy in Figure 4.25. From this figure it can be seen that a magnetic field of ≈14T is required to bend 1 MeV protons in a circle of radius l m.

Figure 4.25. The parameter (BR) for protons moving in a uniform field B, plotted against proton energy.


0.2
0.1
Bρ (T m)
1.0
2.0
3.0
Proton energy (MeV)

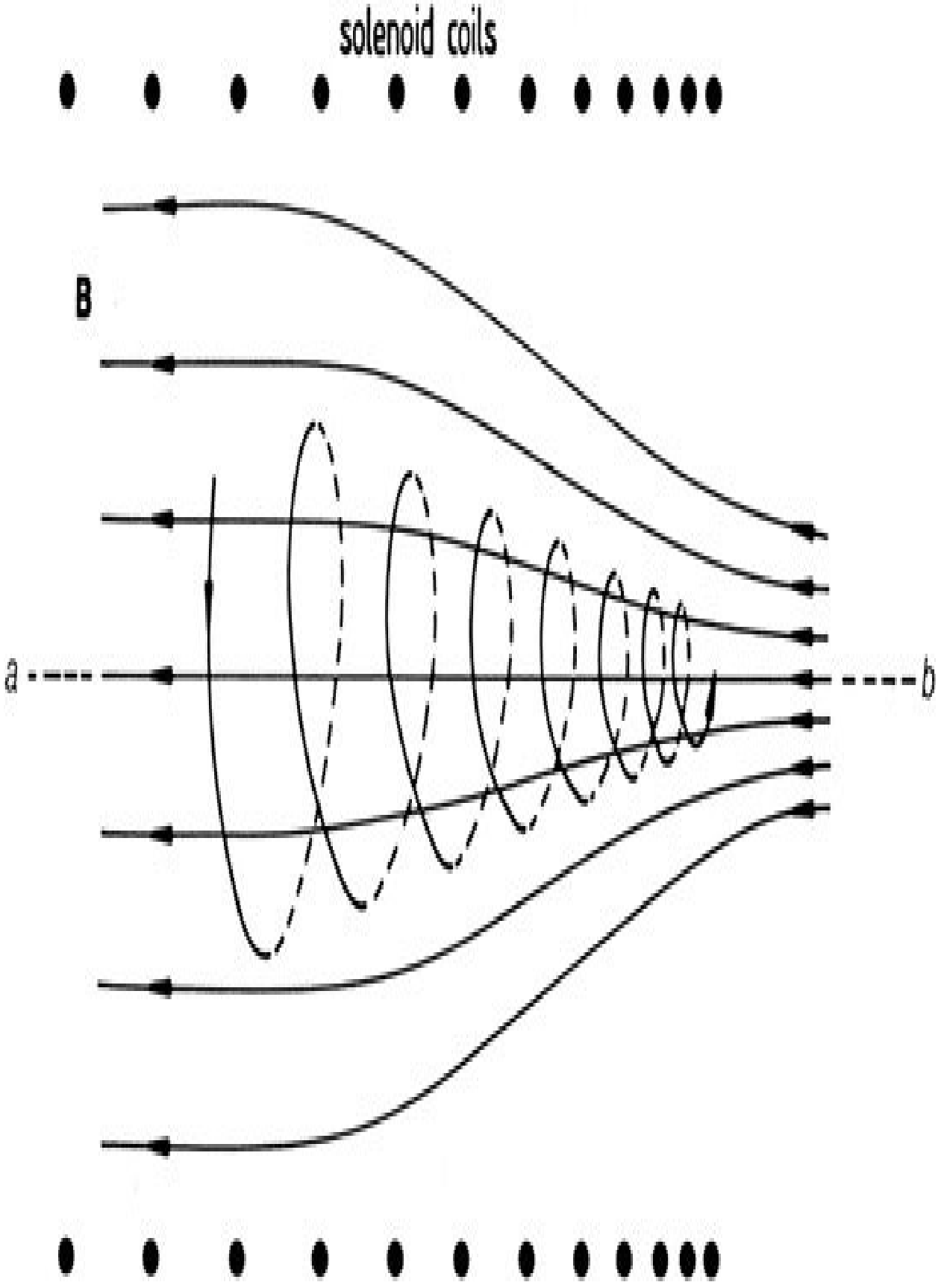
4.7.2 Magnetic mirrors and plasmas

If a charged particle is moving in a uniform field B with its velocity v inclined to the field at an angle other than 90 degrees its motion may be determined by resolving v into two components: parallel to the field B, and $v_\perp$ perpendicular to B. The component of velocity is unaltered by the magnetic force, while the motion in planes perpendicular to B is the circular motion described in section 4.7.1. The resulting path followed by the charged particle is thus a helical trajectory with net translation at speed in the direction of B.

If the magnetic field is produced by a solenoid with more turns per metre at one end than over the remainder of its length the field is non-uniform, being stronger where there are more turns than elsewhere. This is illustrated in Figure 4.26. The non-uniformity introduces into the field a component which is directed towards the axis of the solenoid. If the particle is moving towards the non-uniform field region this component in turn gives a force in the direction opposite to by virtue of the component of velocity in planes perpendicular to the axis. The effect of this force on the overall motion of the particle depends on how fast the ion was moving along axis before entering the non-uniform field region. However, for certain conditions the component of the velocity of the ion may be reduced to zero. The ion will then reverse its direction and traverse the same spiralling path backwards. The non-uniform field has acted to reflect certain ions, i.e. it acts as a magnetic mirror. Since the total magnetic force always acts at right angles to the particle velocity it can not change the particle's energy and if becomes zero $v_\perp$ must increase to keep the energy constant. The actual path of a reflected ion is a helix of decreasing radius with the ion showing down along the axis, coming to a halt and reversing along the same path by which it came, as shown in Figure 4.26.

Figure 4.26. The path of a positively charged particle moving in a magnetic field produced by a non-uniform solenoid. The particle has a component of velocity,

along the axis a–b of the solenoid in the direction from a to b, and this component decreases as the field becomes stronger. Under certain conditions the component may reduce to zero and the particle then retraces its helical path backwards.


solenoid coils
B
a
b

If a similar non-uniform field region is provided at the other end of the solenoid certain particles are trapped in the field and can not escape unless they suffer collisions which upset the requirements for confinement. The solenoid is then acting as a magnetic bottle. Confinement of ions in a magnetic bottle has been used in experiments where the ion density is low but collisions make solenoidal-type bottles inefficient for high ion densities.

The confinement of charged particles within a restricted volume is important for studies of plasmas and their possible uses for energy generation. A plasma is a gas of ionized particles which is electrically neutral. Gas atoms may or may not be completely ionized depending on the temperature of the plasma constituents, defined in the usual way in terms of their average kinetic energies. In sufficiently hot plasmas containing ions of light atoms, the nuclei of the atoms may be moving fast enough to interpenetrate and cause nuclear reactions with release of energy. This is the principle on which fusion reactors, which are presently in the development stage, are based. The practical difficulty is one of confining a sufficiently dense and hot plasma for long enough time for the economic benefit of the energy generated to exceed the economic effort that went into its generation. The confinement problem has proved immense. Prototype devices for power extraction from plasmas use a combination of the internal magnetic fields which arise from the large currents in the plasma and externally applied magnetic fields. The design aim is to shape the external fields and the containing vessel to increase the length of time for which the plasma is stable. The most promising shape, adopted in devices called tokamaks, is one in which the plasma is confined within a toroid.

4.7.3 Magnetic quadrupole lenses

As another example of the motion of charged particles in magnetic fields we will discuss the focusing action of a magnetic quadrupole lens on a beam of moving charged particles.

Consider the idealized situation shown in Figure 4.27, in which all the particles in the beam are moving with the same speed parallel to the z-axis, but are spread out over a limited area in the x–y plane. They enter a region of length over which there is a magnetic field which is designed to focus the beam at the point F, distant f from the field region. If the magnetic field had a perfect focusing action the lens would bring the beam to a point focus at the point F, in a manner similar to that in which a thin convergent glass lens focuses a plane wave at its focal point.

Let us first consider a beam whose profile is a narrow slit in the x direction. A particle moving at a distance xi off the axis of the lens must be deflected through an angle θi which is proportional to the distance xi in order that all particles pass through the focus F. A magnetic quadrupole field within the region between the planes z = 0 and z = l will have this effect on the charged particles. If the particles are positively charged the field lines are similar to those shown in Figure 4.28, which is a view looking from the focus F. In this magnetic quadrupole field the field is given by

(4.59)

where g is a constant. The x-component of the field is proportional to the coordinate y and the y-component to the coordinate x. The field is usually provided by current carrying coils wound around iron pole pieces, and the constant g depends on the current and the distances apart of the pole pieces and their sizes. The length l in Figure 4.27 corresponds to the length of the pole pieces. In practice the field will depend on the coordinate z because of end effects, but we will neglect these, and assume that the field is given by Equation (4.59) for $0 < z < l$, and is equal to zero elsewhere.

Figure 4.27. The focusing action of a quadrupole lens. Particles moving parallel

to the z-axis with non-zero values of the coordinate x are deflected in the magnetic field and are all focused at the point F.

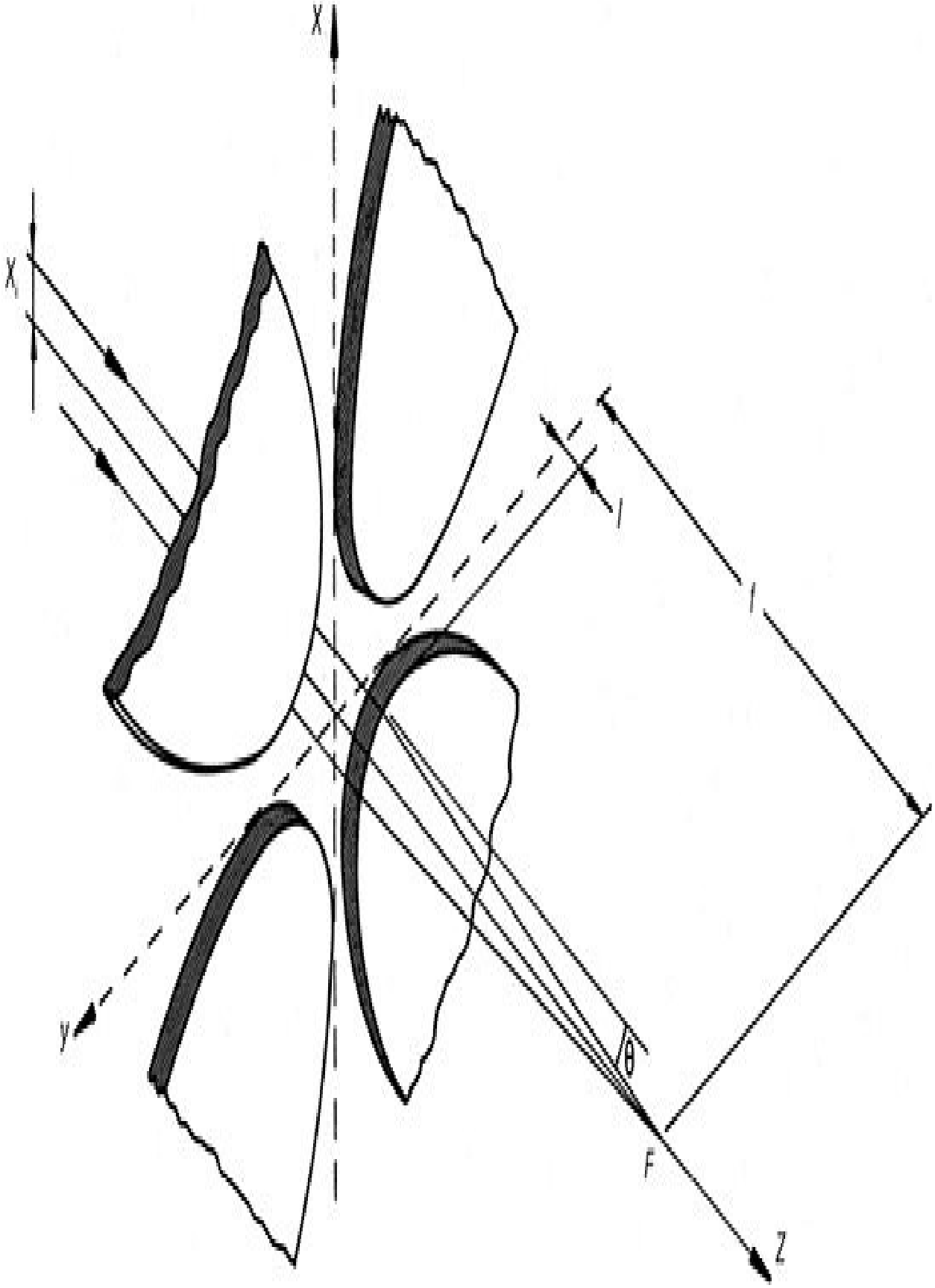
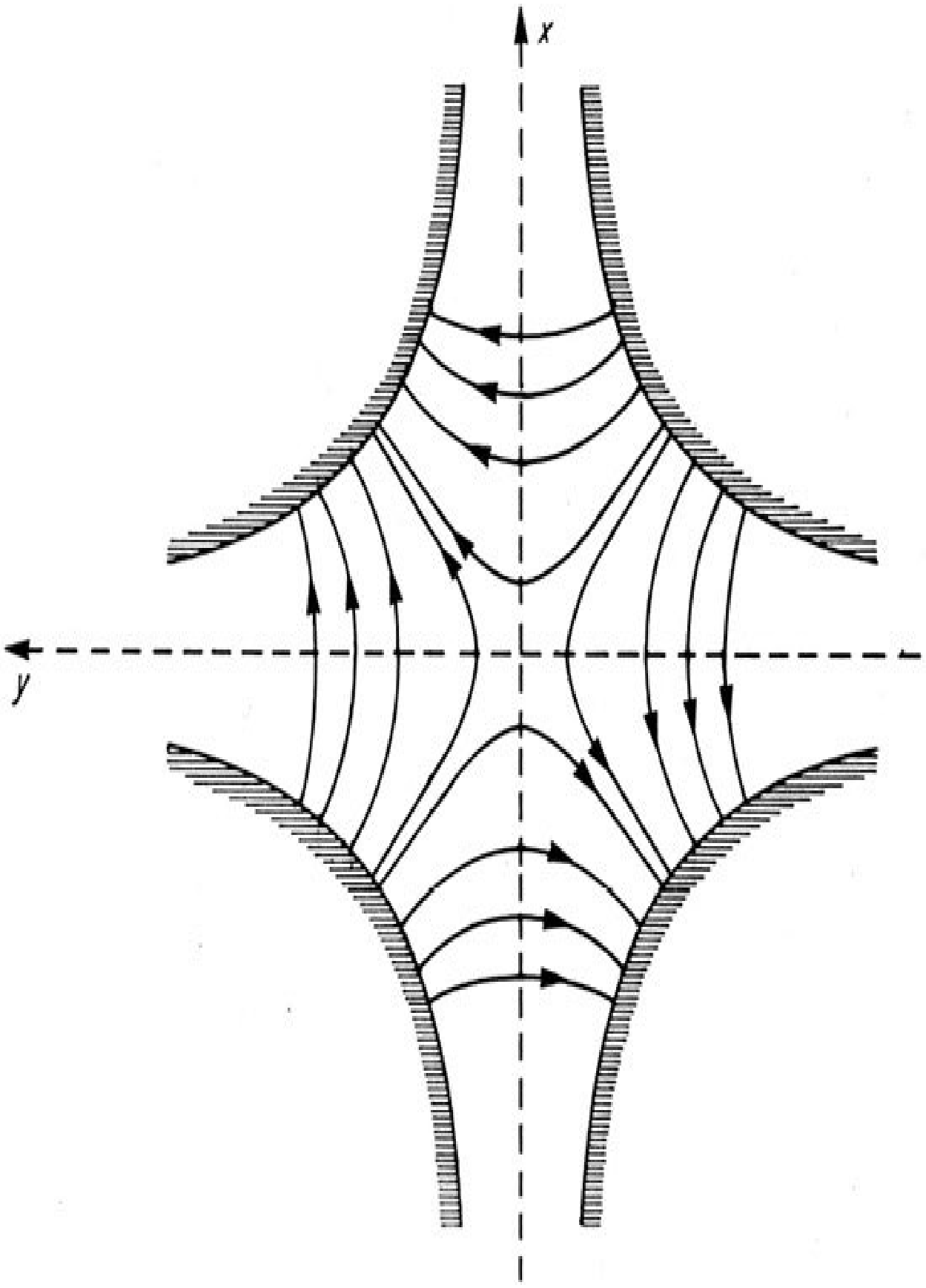
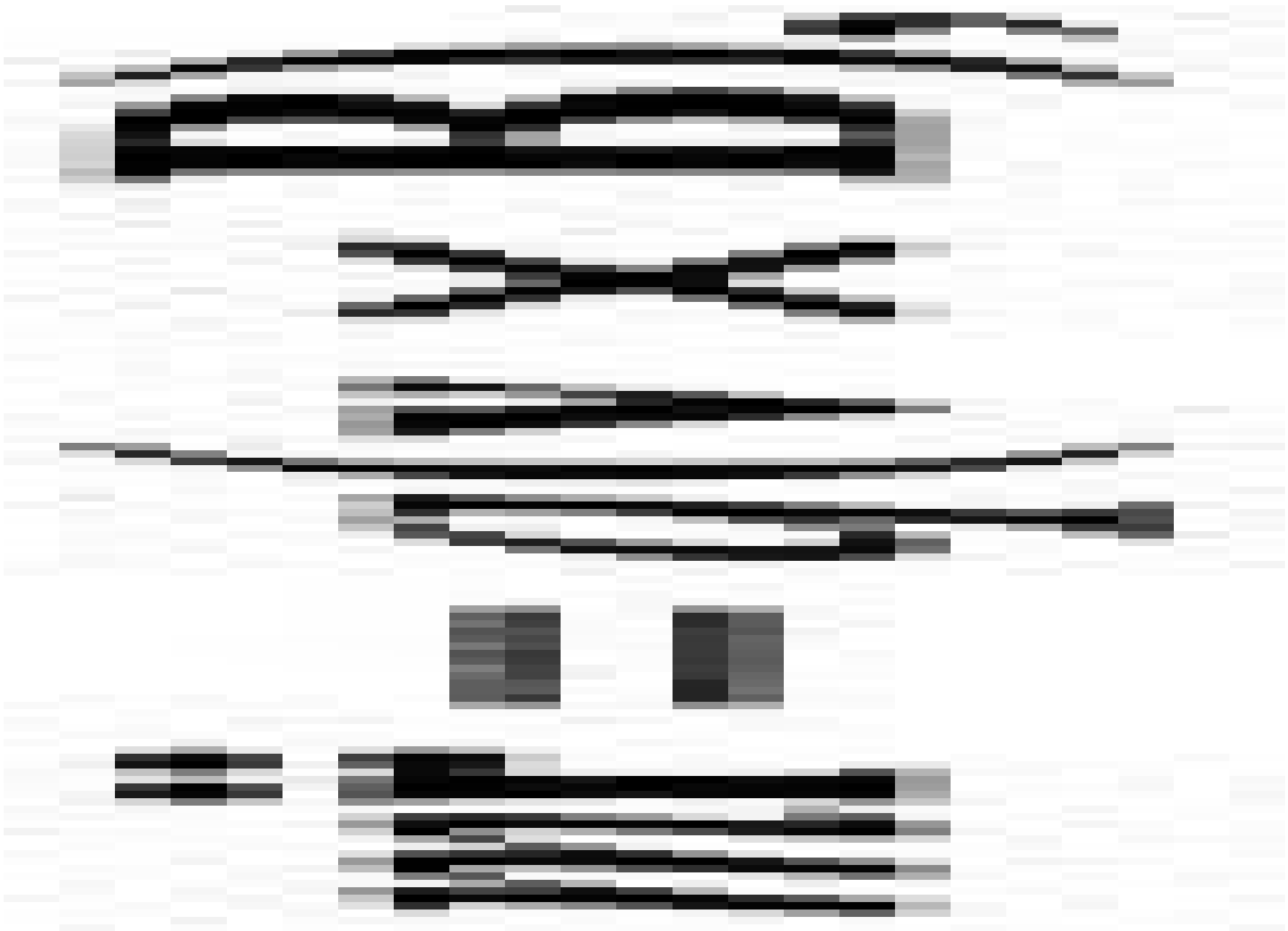

X
X_1
l
l
y
θ
F
Z

Figure 4.28. The field lines of the magnetic quadrupole field which will focus a beam of positively charged particles moving in the z-direction if the initial beam profile is a slit in the x-direction.


X
y

The equation of motion of a particle in the space between the pole pieces is



where m and q are the mass and charge of the particle. If the length l is small compared to the distance f, the velocity is unchanged to a good approximation, and the vector v is constant, with magnitude equal to v, the original speed along the z direction. With this approximation the components of the above equation may be written:

(4.60)

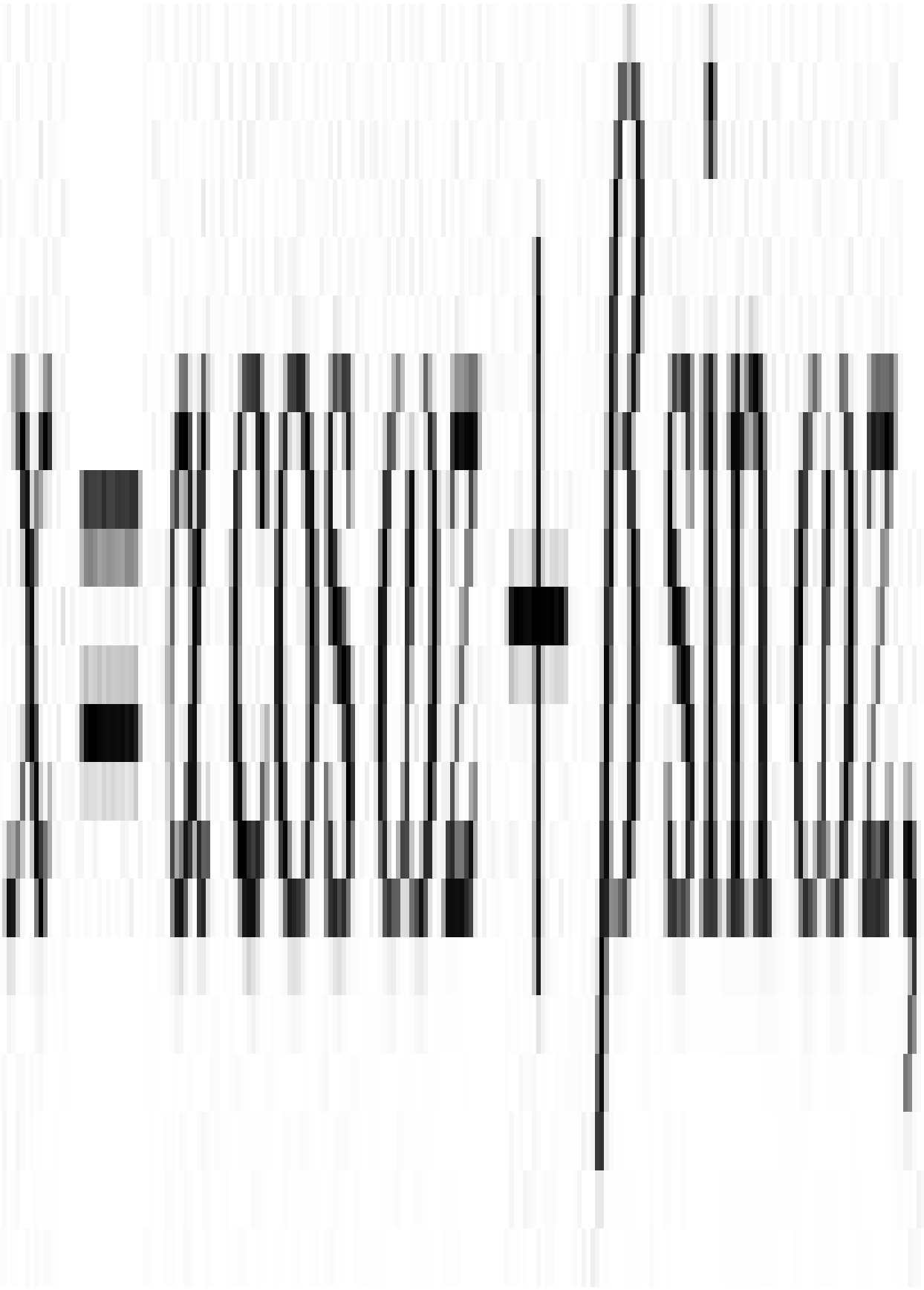
(4.61)

(4.62)

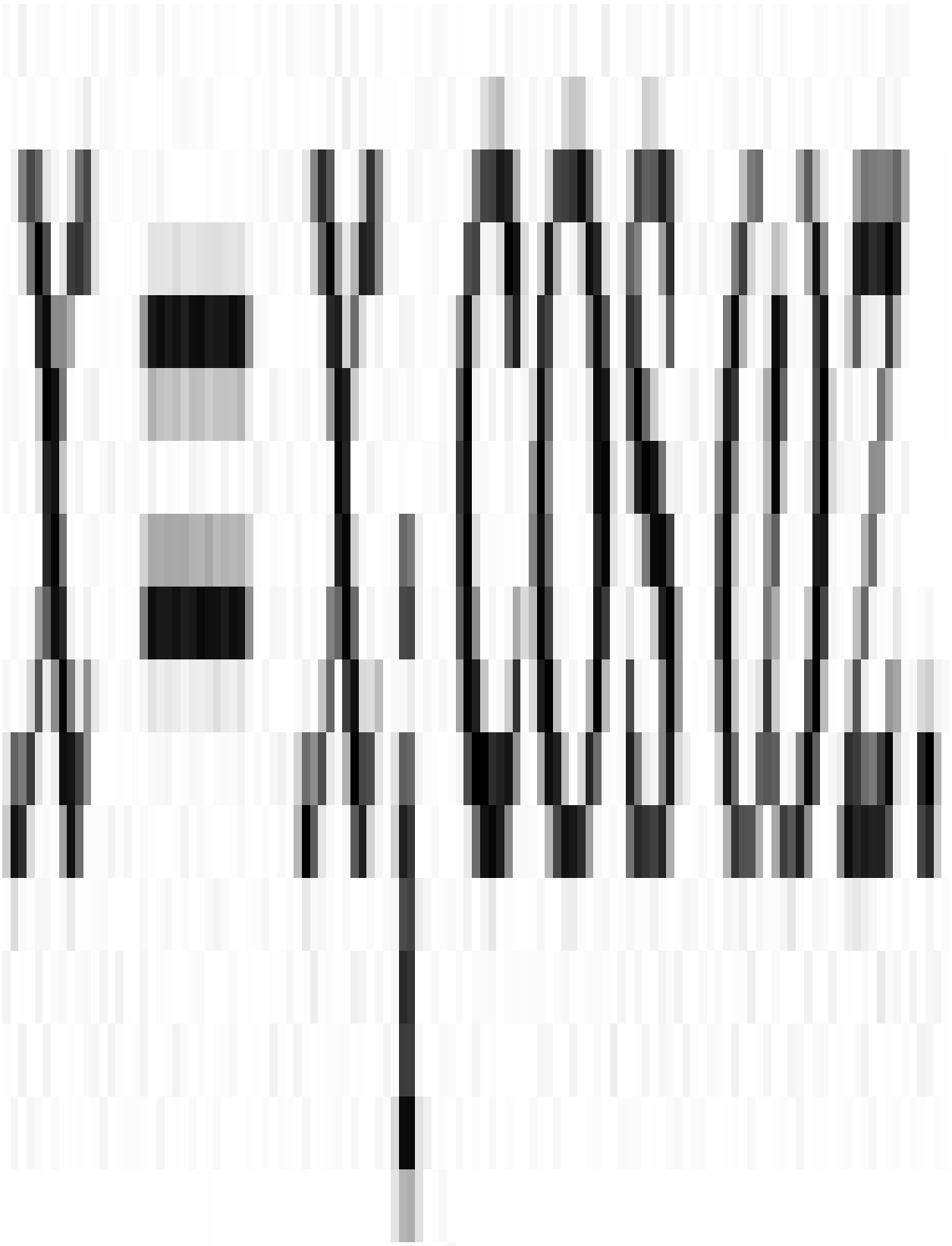
These equations must be solved with the initial conditions that at time t = 0 (the time the particle enters the field region), z = y = 0; dx/dt = dy/dt = 0; and x = xi. Equation (4.60) describes simple harmonic motion in the x direction and its solution is

(4.63)

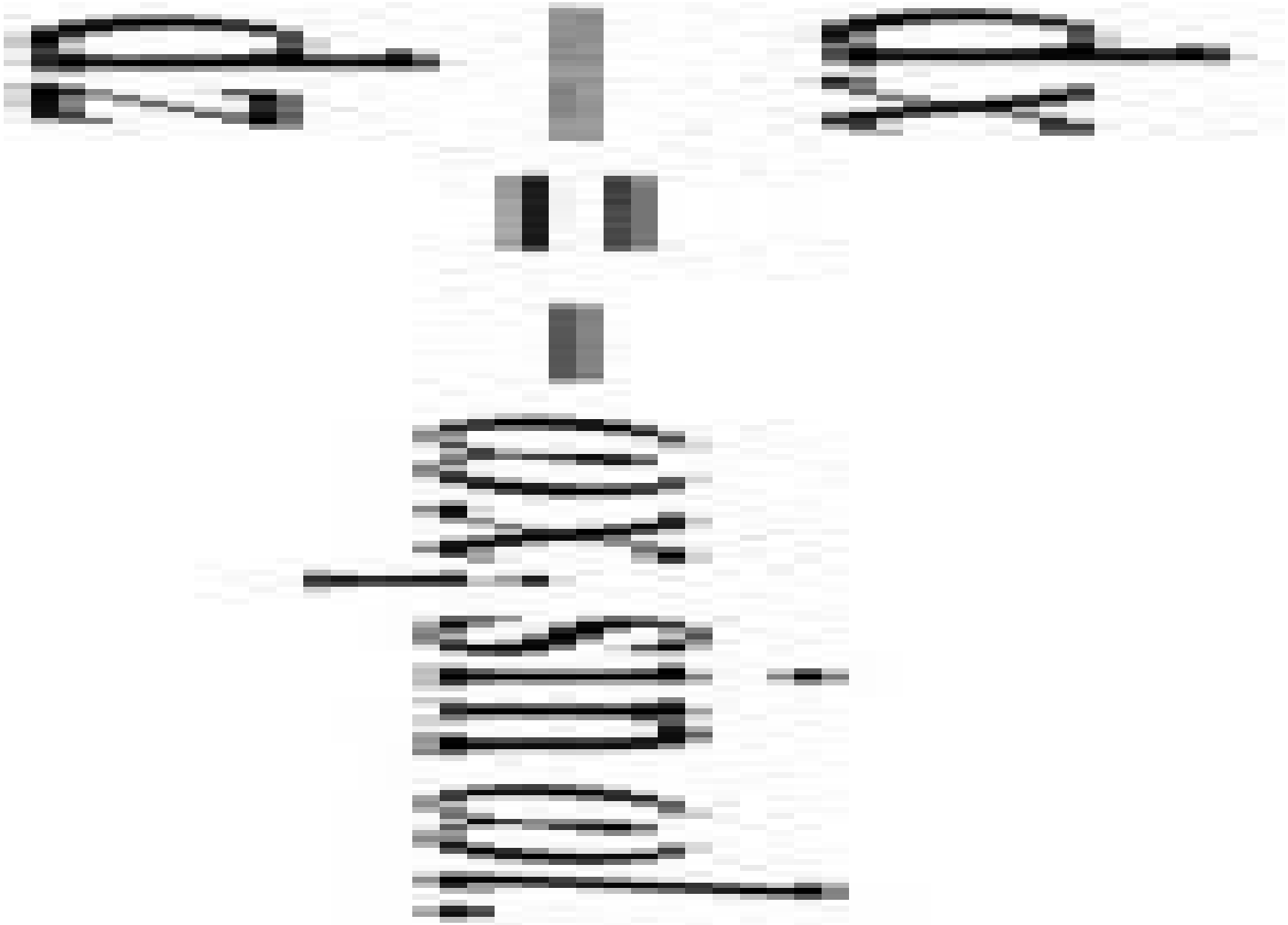
where α and β are constants and The solution to Equation (4.62) is z = vt, and Equation (4.63) may be written, now regarding z as the independent variable, as



where $\omega^2 = gq/mv$. The constants α and β in this equation are determined by the initial conditions to be $\alpha = xi$ and $\beta = 0$, and we have finally



At z = l, when the particle leaves the field,



If ωl 1, which is usually the case and corresponds to the condition used above that f l, dx/dz may be rewritten

(4.64)

The solution to Equation (4.61) shows that the y coordinate is unchanged throughout. Equation (4.64) then tells us that particles initially moving parallel to the z-axis at a distance xi are bent through angles proportional to the distance xi. To a first approximation all the particles in the hypothetical beam we started with pass through the point F on Figure 4.27.

The focal length f is given by

(4.65)

It can be seen that the focusing action of a magnetic quadrupole lens (with fixed dimensions and current) is the same for all particles that have the same value of the quantity (mE/q^2), where m, E and q are mass, energy, and charge of the particles respectively.

If an analysis similar to that above is made for a beam whose profile is a narrow slit in the y direction, it can be shown that the action of the quadrupole field given by Equation (4.59) is defocusing. Particles initially off the axis of the lens in the y direction are bent farther off. Thus for a realistic beam profile the lens described above would make the profile at the focus F into a long slit along the y direction. To give a better focusing action for realistic beams two magnetic

quadrupole lenses are used, one placed very close to the other, the pair constituting a quadrupole doublet. The field in the first lens focuses in one direction, and the field in the second is arranged to focus in a direction at right angles. The combined action is focusing in both directions, rather like the way in which a pair of thin glass converging and diverging lenses can be made to give a converging combination.

The field required in a magnetic quadrupole lens to give it a specific focal length for moving charged particles can be calculated from Equation (4.65). For example if a lens of length 10 cm is required to have a focal length of 3 m for 50 MeV protons, the field gradient g, given by

has to be about 3.3 T m−1.

PROBLEMS 4

4.1 Estimate the mean velocity at which oxygen molecules carrying a net charge (− e) would move between two parallel plate electrodes 10cm apart and having a potential difference of 100 V between them. Estimate the negative current if each electrode is of area 10 cm^2. The oxygen between the plates is at S.T.P.; at S.T.P. the r.m.s. speed of the molecules is 3×10^4 cm s−1 and the mean free path between collisions is 10-5 cm. One part in 10^{11} of the oxygen molecules is ionized.

4.2 Two spherical buoys float half submerged in a calm deep sea. Their radius a is very much smaller than their separation b. Calculate the resistance between them if the resistivity of the water is ρ.

4.3 A small sphere is uniformly charged throughout its volume, and is rotating with constant angular velocity. Determine its magnetic moment in terms of the total charge Q on the sphere, the angular momentum L of the sphere, and its mass M.

4.4 Two very long thin wires carrying equal and opposite currents of 10^3 A are placed parallel to the x-axis at y = 0 and z = ± 0.5 cm. Calculate the magnetic field B in the x–y plane and determine its maximum gradient.

4.5 A long sheet of conductor of thickness 1 cm and height 20 cm carries a current of 10^4 A distributed uniformly within it. Calculate the magnetic field along a line ab perpendicular to the surface cutting the sheet half way up as shown in Figure 4.29. Consider distances from the sheet small compared with 20cm.

What is the field inside the sheet? What is the force on an electron travelling near the speed of light parallel to the sheet?

4.6 A long straight wire of circular cross-section carries a current I uniformly distributed. What is the magnetic field inside the wire?

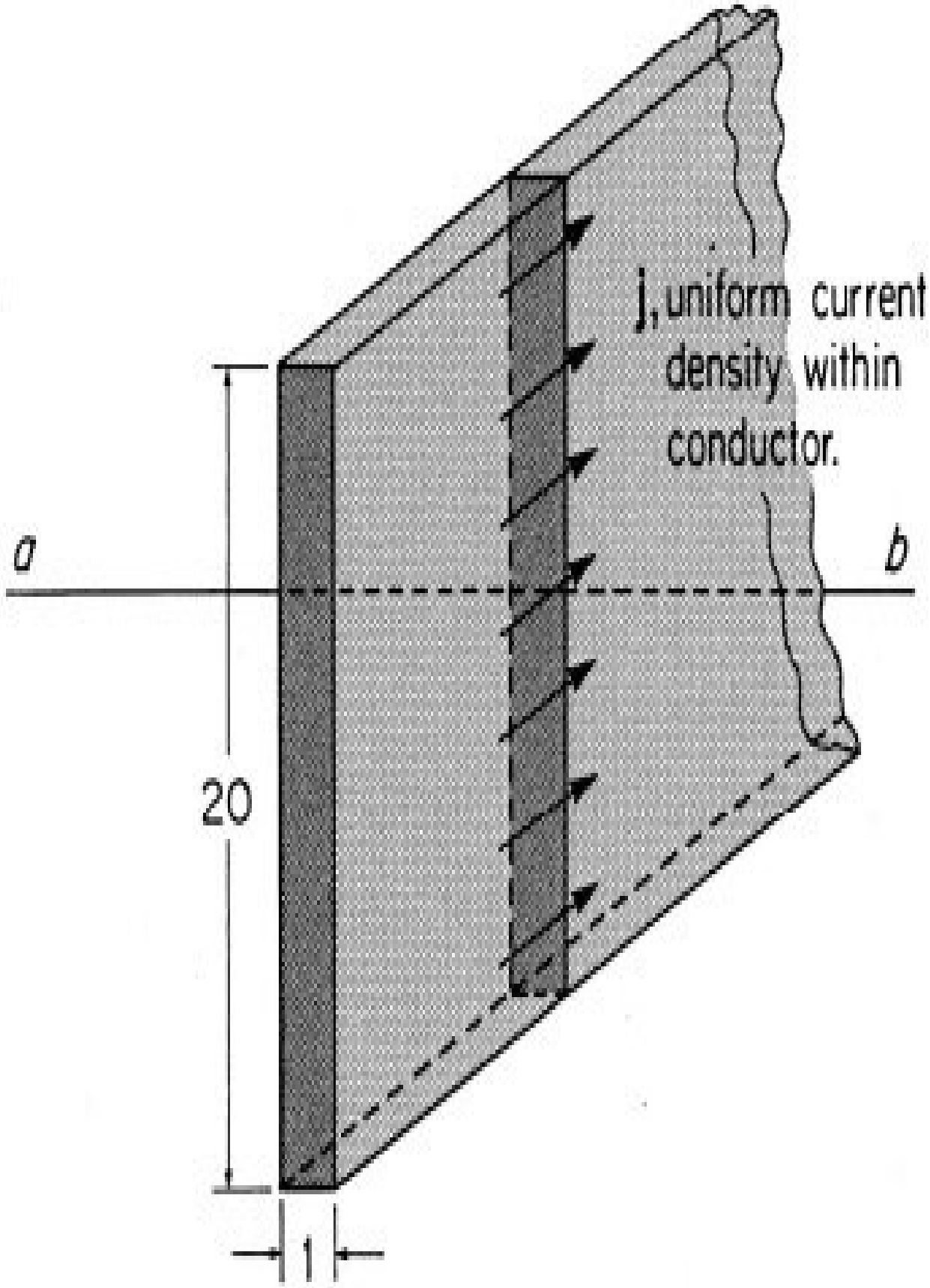
4.7 A cable of circular cross-section and diameter 2 cm has a long cylindrical hole of diameter 1 mm drilled in it parallel to the cable axis. The distance

between the axis of the hole and cable axis is 5 mm. If the cable has a uniform steady current density of 10^5 A m^{-2} flowing in it, calculate the magnetic field

(i) at the centre of the cable

(ii) at the centre of the hole.

Figure 4.29.


j, uniform current
density within
conductor.
a
b
20
1

4.8 Using Equation (4.34) for the field B outside a long wire carrying a current verify that curl B = 0 outside the wire. Determine curl B inside the wire.

4.9 A flat coil having 50 turns of radius 30 cm is in series with another flat coil of 100 turns each of radius 2 cm, the coils being coaxial and 10 cm apart. Estimate the force between the coils when a current of 2 A is passing.

4.10 The magnetic field of the Earth is found to fall off according to an inverse cube law at several Earth radii. The field can thus be roughly described in terms of a magnetic dipole at the centre of the Earth. If the field strength is 10^{-4} G at 15 Earth radii in the equatorial plane what is the magnetic dipole moment of the Earth? (Radius of Earth 6000 km.)

4.11 A particle of mass m, charge q, moving with velocity v enters a region where there is a uniform magnetic field B. The vectors v and B make an angle θ with each other. Determine the motion of the particle.

4.12 An electron moving in outer space has a component of velocity 0.01c in the direction of a magnetic field of 10^{-5} gauss. How many revolutions does it make whilst travelling between two points one light year apart along a line of magnetic field? (Treat the problem non-relativistically.)

4.13 Show that the torque, given by Equation (4.51), on a coil carrying a current in a magnetic field reduces to Equation (4.25) as the area of the coil, δS, becomes very small, such that the coil corresponds to a dipole of moment $m = I\delta S$.

4.14 In regions of space where the current density is zero, the magnetic field can be obtained by taking the negative gradient of a scalar function ψ, in the same way as the electrostatic field can be obtained from the electrostatic potential ϕ. The potentials ψ and ψ both satisfy Laplace's equation (3.4). Show that in cylindrical polar coordinates (r, θ, z) the function $\psi = Ar^n \sin(n\theta)$, with A and n constants, is a solution of Laplace's equation and describes a field B which has magnitude Anr^{n-1} independently of z and θ.

A hydrogen atom, with magnetic moment 9.27×10^{-24} A m^2, is in a field of the above form with A = 100 T m^{-2} and n = 3. It is situated 1 cm from the origin at a point where the direction of the field is along the radius vector, e.g. at $\phi = \pi/2$, and has its dipole moment oriented along the direction of the field. What is the

force on the atom?

4.15 Particles with velocity v and magnetic rigidity mv/q equal to 16 T m enter a dipole magnet which provides a uniform magnetic field B perpendicular to v. The product of field and magnet length is often called the bending power of the magnet. The particles are deflected through a small angle θ, which is the angle between the incoming and outgoing directions. If the bending power of the magnet is 2 T m estimate the angle of deflection.

4.16 Use the Biot-Savart law and Equation (4.53) to calculate the force between two thin current-carrying coils each consisting of a single loop of wire. Use Equation (4.50) to give an equivalent expression for the force. (Although the two expressions look different it can be shown that they are the same.)

*This is not the total magnetic energy associated with the entire physical system, which consists of the current I in the coil and the sources that give rise to the external field. The total energy is discussed in Chapter 6.

*Equation (4.58) has been derived non-relativistically. It is valid for any speed of the moving particle however, if one uses the relativistic mass given by $m = mo(1 - V^2/C^2)^{-1/2}$, where mo is the rest mass of the particle.

CHAPTER 5

Magnetic materials

Up to now we have been discussing magnetic fields in free space, or in air. In air the distances apart of the molecules are relatively large, and the influence of the air molecules on the magnetic fields is very small. We have seen how a steady current gives rise to a steady magnetic field, and have obtained equations describing the field's behaviour. Suppose that we put a sample of material into a magnetic field. How is the field changed and what are the new field equations? How do we describe what happens to the material? What new phenomena arise that can be measured and used to learn something about the material's microscopic structure? These questions are discussed in this chapter.

5.1 MAGNETIZATION

If a specimen is placed in a magnetic field, the field in the vicinity of the sample is changed. How much it is changed depends on the material and on the shape of the sample, but for most specimens it is not altered very much—typically only by about one part in 10^5. For certain materials, however, the field at some points nearby can be increased by a factor of one hundred or so over its value in the absence of the specimen. Such materials are called ferromagnetic. Nonferromagnetic materials, which produce only a small change in field, can be either diamagnetic or paramagnetic. If a diamagnetic rod is inserted into a solenoid that carries a current, the field near the end of the solenoid is decreased very slightly. If a paramagnetic rod is inserted the field is increased slightly. Glass and copper are examples of diamagnetic materials; oxygen and titanium are paramagnetic; iron, cobalt and nickel, and certain alloys of these metals, are ferromagnetic.

The most easily observed characteristics of the different materials are their behaviours in strong non-uniform fields. A ferromagnetic material is strongly pulled into the field. A diamagnetic material is repelled towards field-free regions, but so weakly that sensitive apparatus is required to measure the repulsive force. A paramagnetic material is pulled into the magnet. Although the attraction is usually weak, it can sometimes be strong enough to observe in a simple way. If a flask full of liquid oxygen is placed near the gap in the yoke of a strong electromagnet the liquid may flow out of the flask and into the gap between the pole pieces.

The explanation of these effects lies in the interaction of microscopic magnetic dipole moments with the fields in which the materials are placed. Let us restrict our attention to liquids and gases, in which the atoms and molecules move independently. These atoms or molecules may have permanent magnetic dipole moments. If they do have permanent moments, the material is paramagnetic; otherwise it is diamagnetic. If paramagnetic molecules are placed in a magnetic

field there is a net alignment of the dipoles in the field direction. This is exactly analogous to the way in which polar molecules on average acquire an effective electric dipole moment in the direction of an applied electric field, as discussed in section 2.2.2. In a non-uniform magnetic field a paramagnetic material will thus experience a force in the direction of increasing field.

All atoms and molecules acquire an induced magnetic dipole moment when placed in a magnetic field. Induced magnetic dipole moments, unlike induced electric dipole moments, are in a direction opposite to the field in the specimen. Thus, if the molecules of a substance have no permanent moments, i.e. the substance is diamagnetic, a sample inserted into a non-uniform magnetic field will experience a force in the direction of decreasing field strength.

On the atomic scale atoms or molecules in a magnetic field are said to be magnetized. The microscopic magnetic dipole moments can vary from atom to atom but can be averaged over volumes which contain very many atoms to give a function which is smoothly and slowly varying with position on a macroscopic scale. This function is called the magnetization of the medium, and is given the symbol M. It is defined as the magnetic moment per unit volume, and is given by

(5.1)

where N is the number of molecules per unit volume, and $\langle m \rangle$ is their average dipole moment in the direction of the field. As in electrostatics we consider only isotropic materials, for which the magnetization M can be simply related to the field. The dimensions of magnetization are [current] × [length]−1, and it is measured in amperes per metre. As in the electrical case, magnetized atoms produce magnetic fields that fall off more rapidly with distance than the dipole field. These higher order fields need to be described by more complex atomic properties than the average effective dipole moment $\langle m \rangle$. However, they make essentially no contribution to the macroscopic magnetic field, and when

discussing this field it is safe to ignore the higher order microscopic fields completely. Thus, the macroscopic effects of atomic magnetization can be adequately represented by the magnetization M.

Before we discuss the consequences of magnetization on a macroscopic scale we will use simple atomic models which approximate the behaviour of electrons in atoms in order to gain insight into the origin of diamagnetism and paramagnetism. There are two chief contributions to the magnetic dipole moment of an atom or molecule. The first arises from the orbital motions of the electrons, which correspond to current loops. The second arises because electrons themselves have permanent (intrinsic) magnetic dipole moments. These intrinsic moments have a similar magnitude to those due to the orbital motions. The magnetic dipole moments are closely related to the angular momenta of the electrons; the orbital moment to the orbital angular momentum, and the intrinsic moment to an intrinsic angular momentum possessed by electrons. This latter angular momentum, the electron spin, can for many purposes be regarded as due to the electron spinning around an axis through its centre, like a top. This picture should be treated with caution, however, since all experiments performed so far show that the size of the electron is less than 10^{-18} m, and its properties are consistent with it being a point-like object. The proper description of angular momentum in atomic systems requires the use of quantum mechanics and is outside our scope. We need only mention one result pertinent to the present discussion. This is that the component of the spin angular momentum of an electron, measured in any specified direction, can only have the values $\pm \hbar/2$, where $\hbar$ is Planck's constant divided by 2π. Compare this with the assertion in section 4.4.3 that the component of the orbital angular momentum of an electron in a specified direction can take only the value zero or an integral multiple of $\hbar$.

If the electrons in an atom or molecule have zero resultant angular momentum that atom or molecule has no permanent magnetic dipole moment*. A liquid or gas consisting of such atoms or molecules is diamagnetic. If there is a resultant electronic angular momentum there is a permanent moment, and the liquid or gas is paramagnetic. In the solid state, where the atoms or molecules do not move independently, and where a particular electron is not necessarily attached

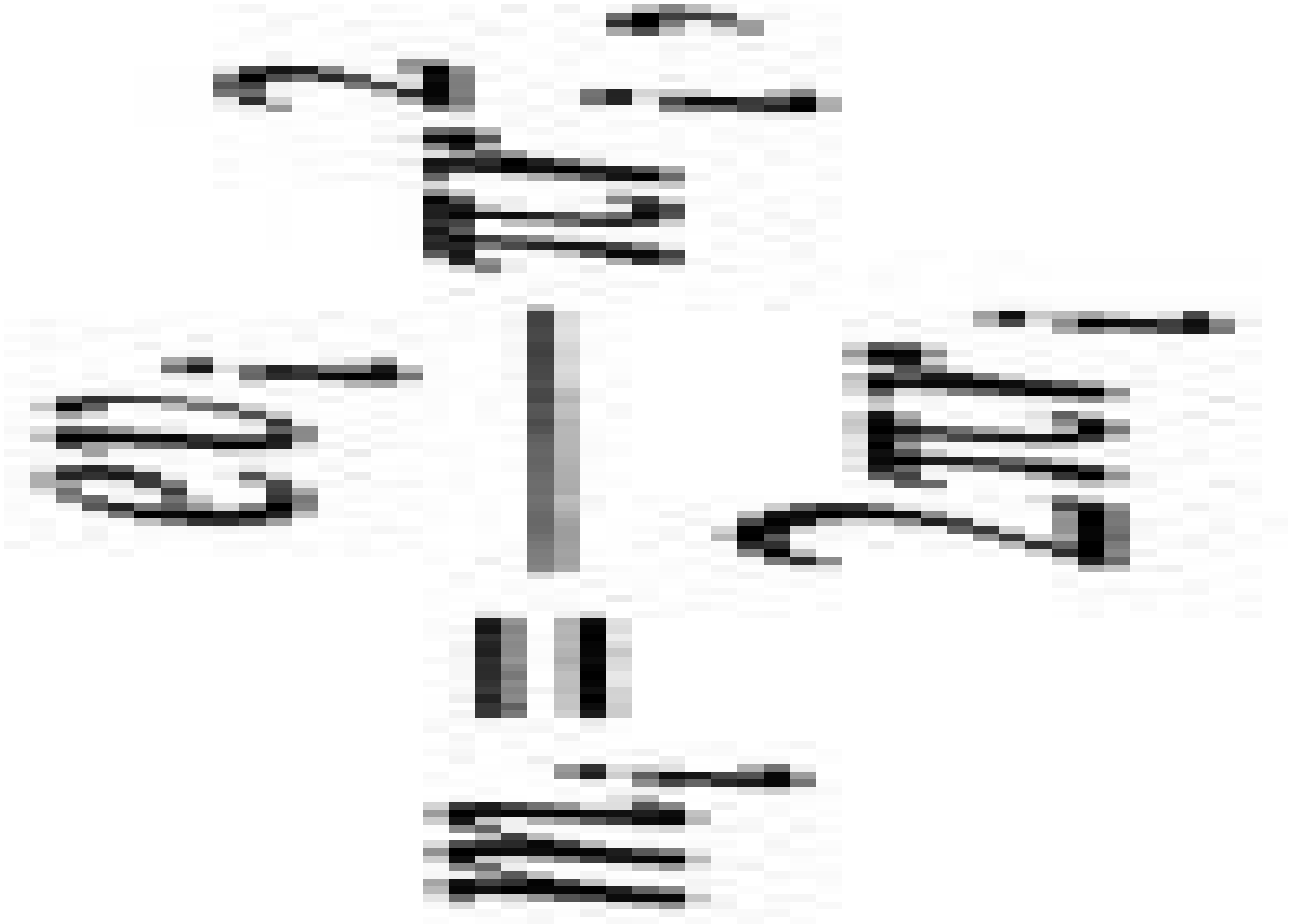
to a single molecule, the situation is more complicated. In a metal for example there are contributions to the magnetization from the conduction electrons and there may be contributions from the ions fixed in the lattice. These contributions vary with temperature and with field strength, and usually result in a small effect which can be either paramagnetic or diamagnetic. Ferromagnetic materials, however, may show large magnetizations. In certain metals, for example, there is an interaction between electrons on the lattice ions and the conduction electrons tending to align all of the latter in the same direction. Complete alignment can occur in large magnetic fields resulting in a very large enhancement of the total field.

5.1.1 Diamagnetism

In this section we consider the origin of diamagnetism. We will use a classical model to describe how atoms acquire induced magnetic dipole moments in a direction opposite to an applied magnetic field.

We first use the simple atomic model outlined in section 4.4.3 to relate the magnetic moment of the orbital motion of the electrons in an atom to their total orbital angular momentum. Consider an electron, which we label i, moving in a circle of radius ri about a nucleus. If the electron has speed vi, the current around the circle is in the direction opposite to the electron’s motion (since the charge on the electron is -e) and has size

The magnetic moment of the.current loop has magnitude



and is in the direction opposite to the direction of the electron’s angular momentum. The magnitude of the magnetic moment can be rewritten in terms of the size of the angular momentum $L_i = m_e\omega_i$ of electron i, where m_e is the mass of the electron and ω_i is its angular speed. We obtain

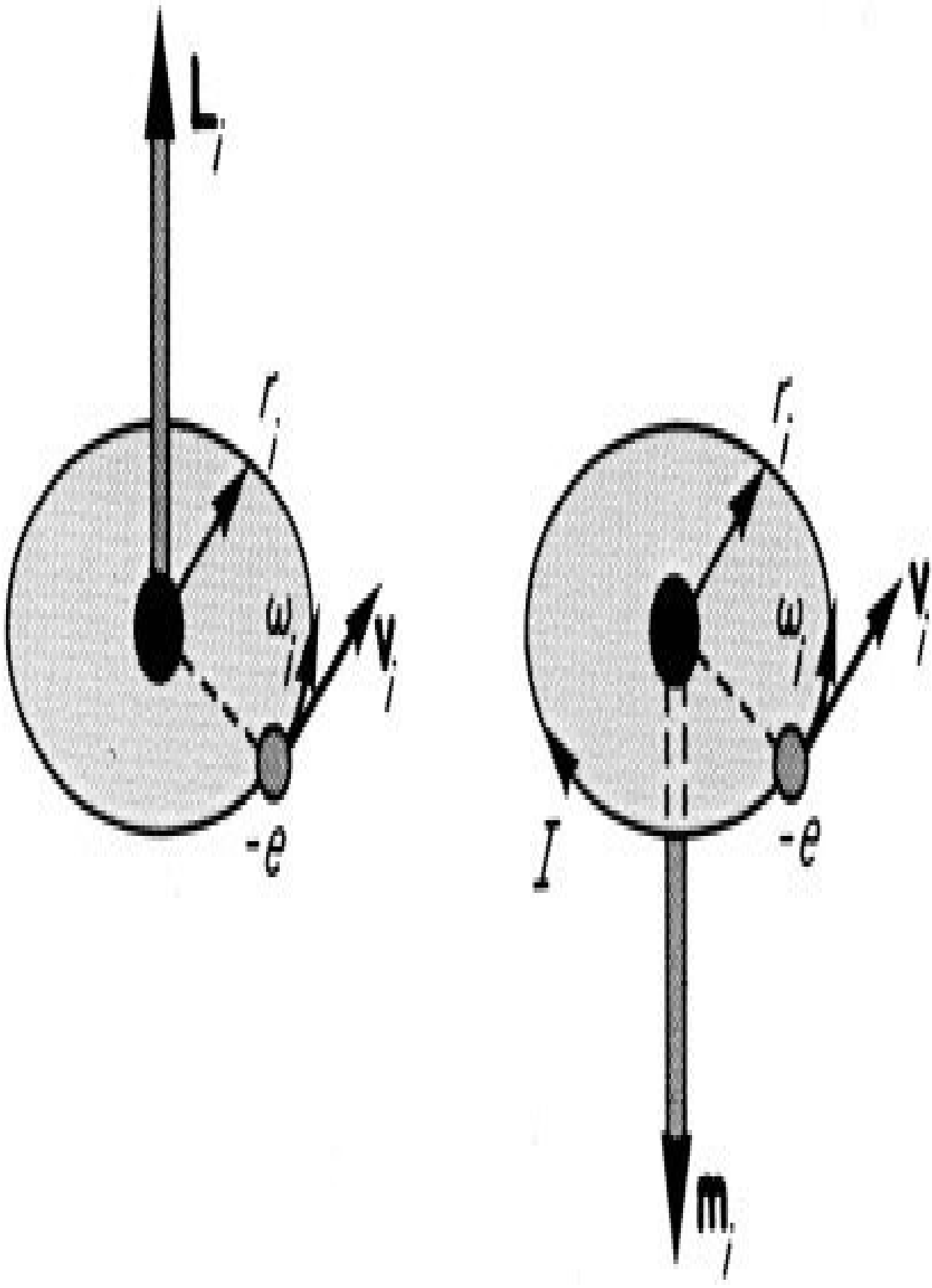
The direction of the vector mi is in the opposite direction to that of the angular momentum Li, as shown in Figure 5.1, hence

(5.2)

The total magnetic dipole moment resulting from the orbital motion of all the electrons in the atom is the vector sum

(5.3)

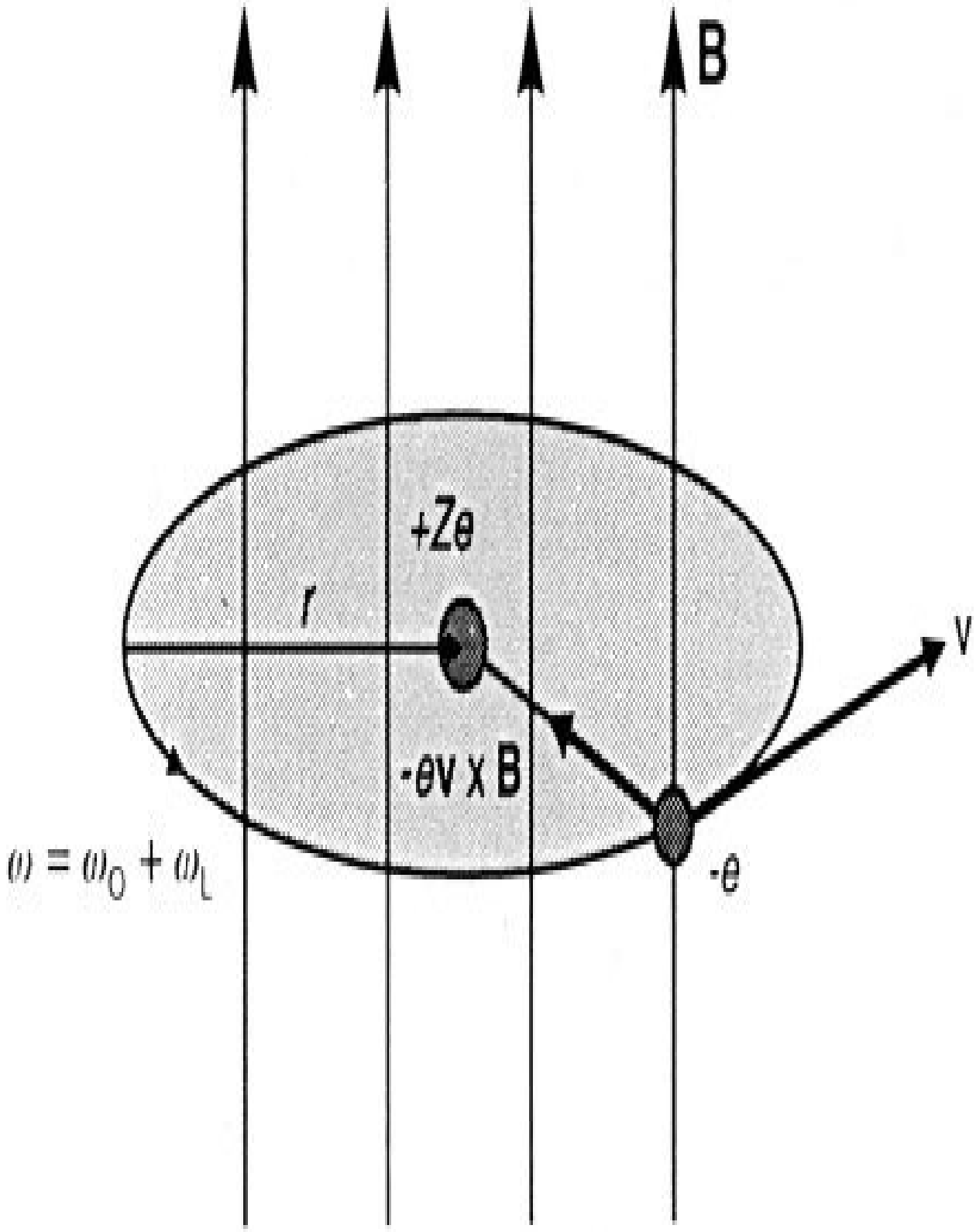
Figure 5.1. The angular momentum and magnetic moment of an electron moving in a circular orbit. The current is in the opposite direction to the motion of the negatively charged electron, and the orbital magnetic moment mi is opposite to the orbital angular momentum Li.

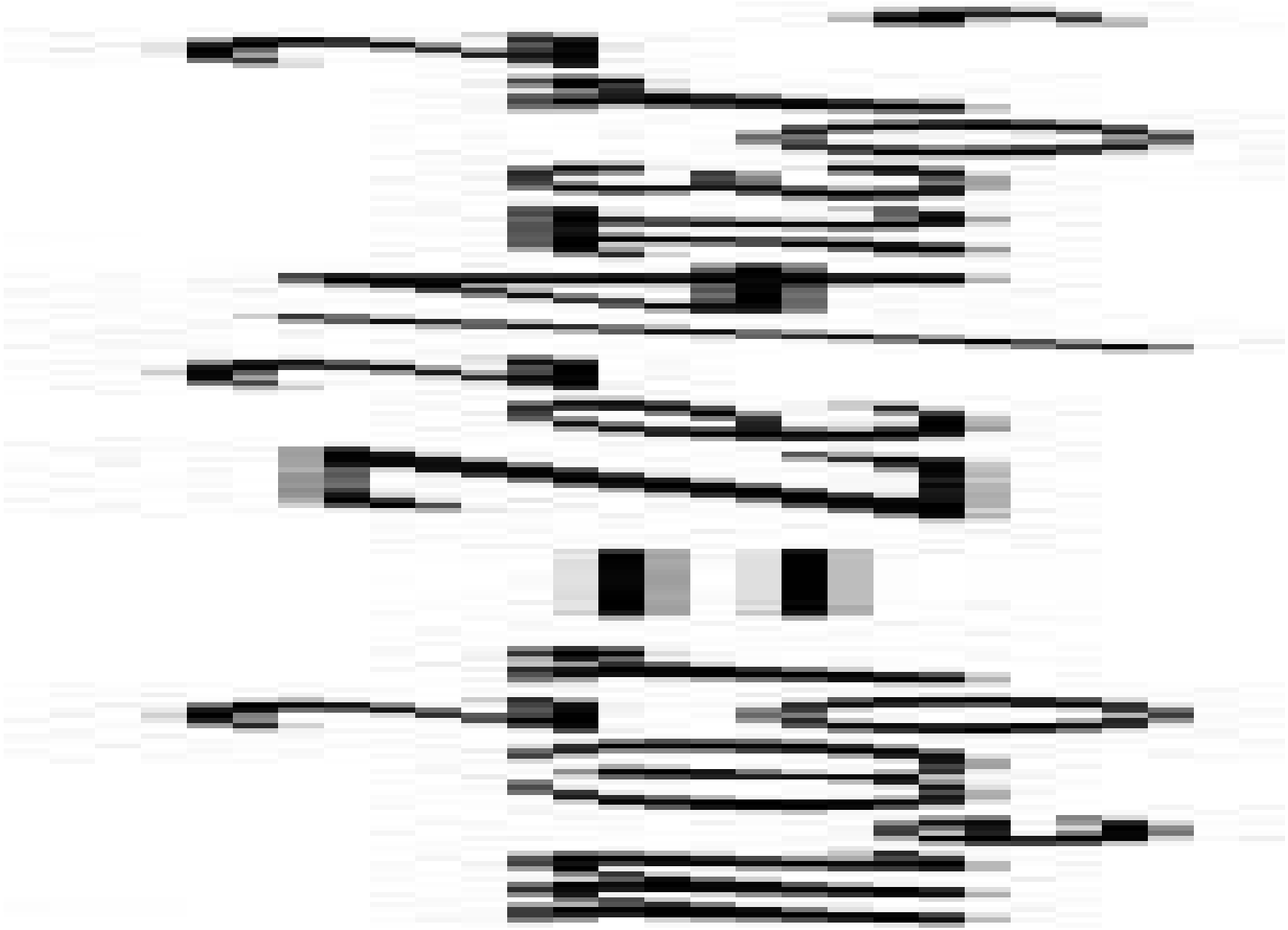

L_i
r_i
ω_i
v_i
$-e$
r_i
ω_i
v_i
I
$-e$
m_i

If the resultant orbital angular momentum is zero, and if the resultant intrinsic angular momentum of the electrons is zero, the atom has no permanent dipole moment and is diamagnetic. The magnetic effects of a material which consists only of diamagnetic atomic particles are entirely due to induced magnetic moments.

Now suppose we apply a magnetic field to a diamagnetic atom. The resultant intrinsic angular momentum of the electrons remains zero but the orbital motions of the electrons are slightly changed. Consider the orbit shown in Figure 5.2 in which an electron of mass me is moving with angular speed ω0 in a circle of radius r. The central acceleration is , where

Figure 5.2. The effect of a magnetic field on the circular orbit of an electron whose orbital angular momentum is in the same direction as the field. The angular speed ω0 is increased by the amount ωL

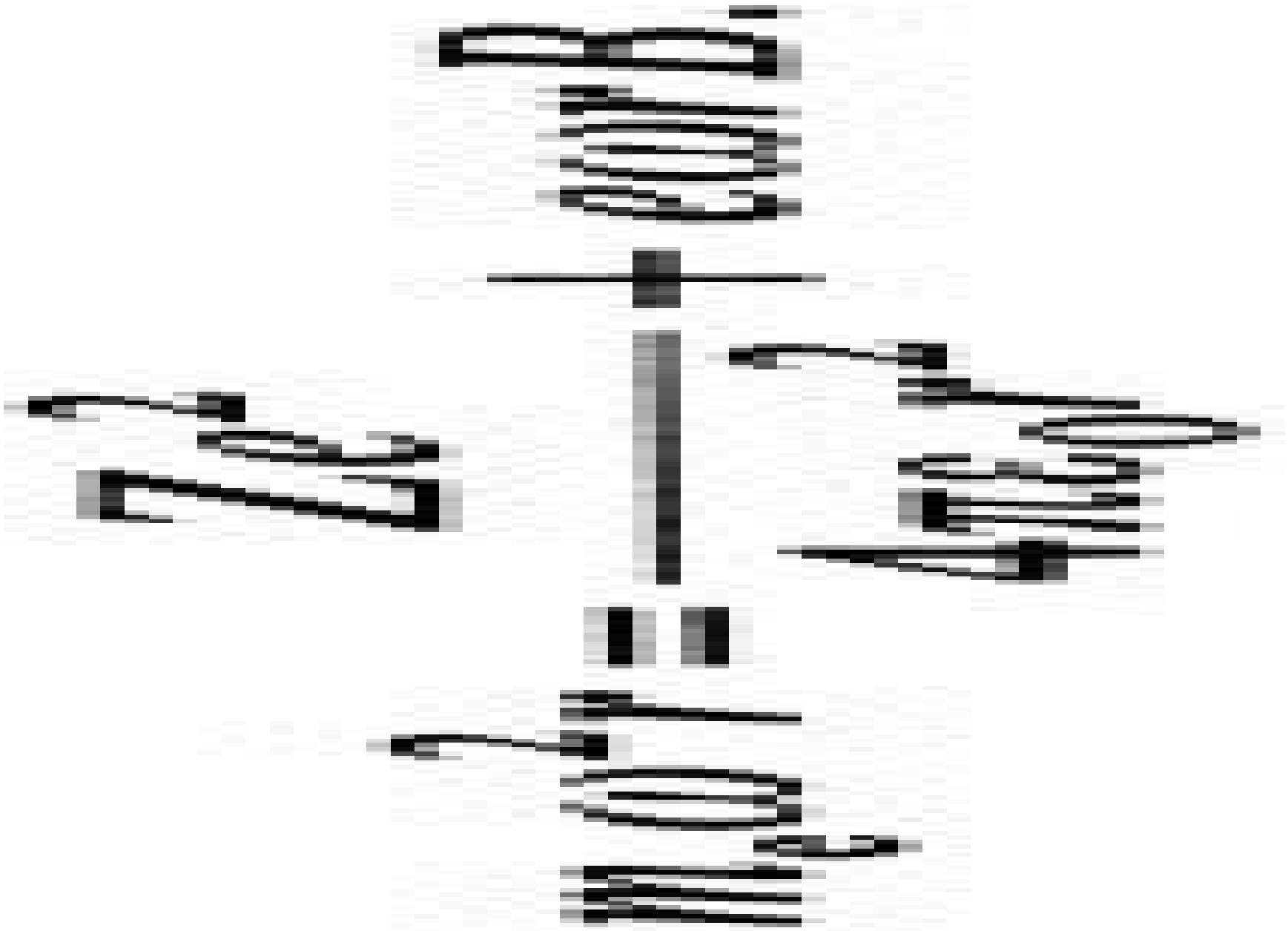

B
+Zθ
r
V
-θV x B
ω = ω0 + ωL
-e



and hence

(5.4)

When a field B is switched on, perpendicular to the plane of the orbit, and in the direction shown, the force on the electron is increased by an amount $-ev \times B$, where v is the velocity of the electron. This extra force is very small compared with the atomic forces holding the electron in the atom, and as a first approximation we will assume that the electron stays in its original orbit after the field is switched on. The angular speed of the electron in the orbit changes to the value ω, where



Solving this quadratic equation for the angular velocity ω gives

(5.5)

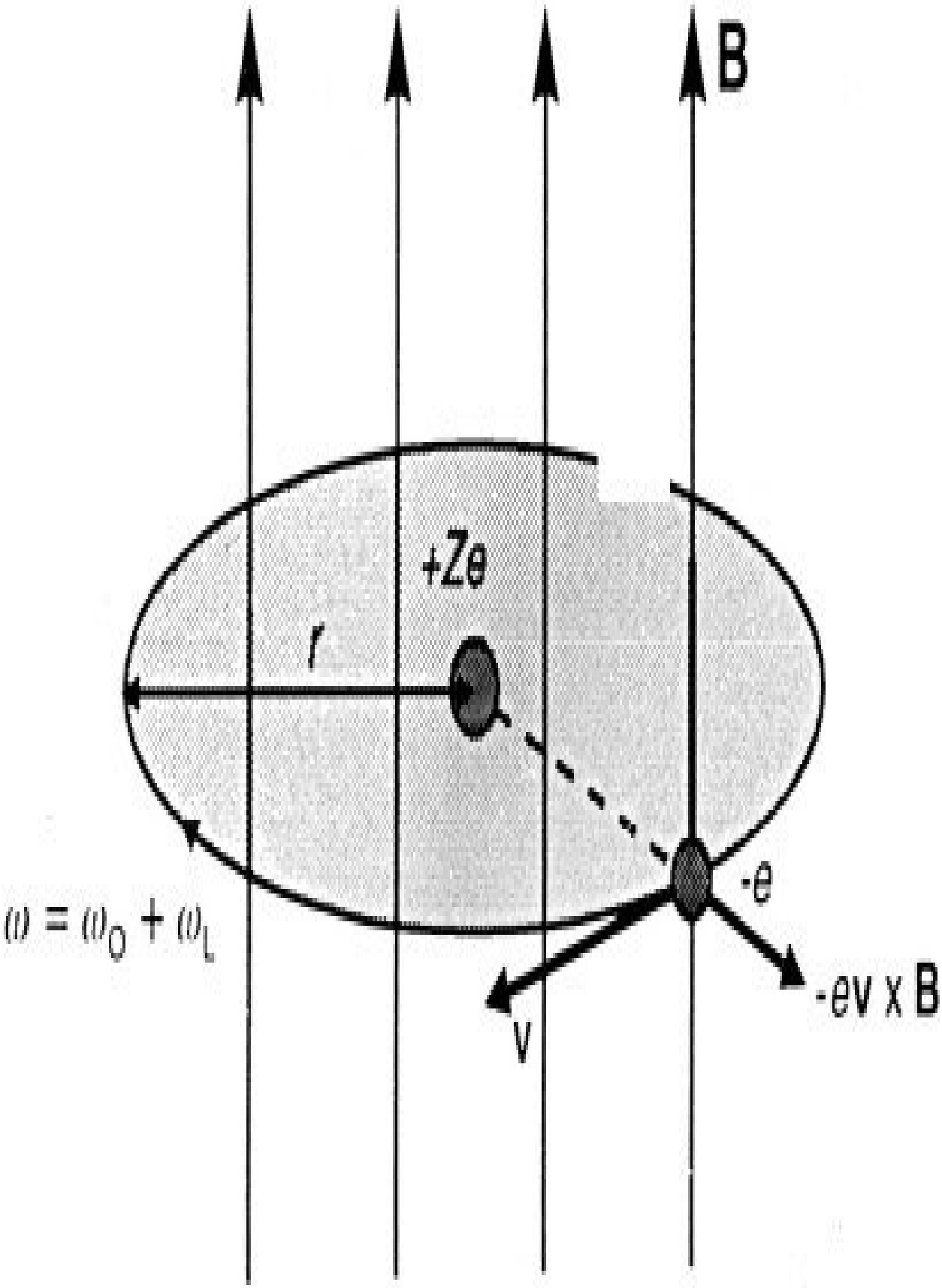
if the condition

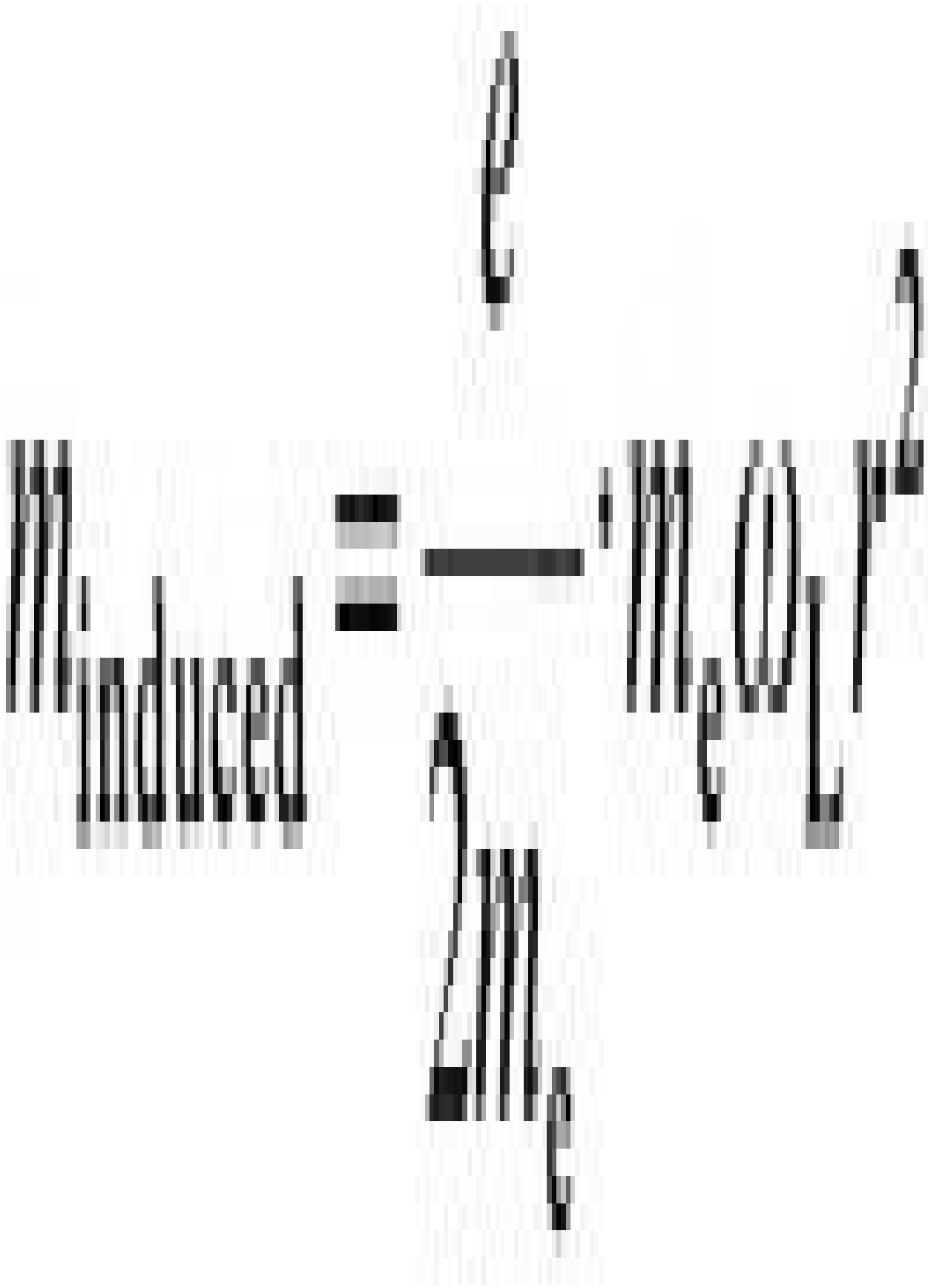
is satisfied. For all normal fields this condition is met. The result (5.5) tells us that in the presence of the field the angular velocity of the electron is increased by an amount (eB/2me), a quantity known as the Larmor angular frequency and often written ω_L. The increase in angular frequency gives the electron shown in Figure 5.2 an increase in angular momentum equal to $m_e\omega_L r^2$.

Now consider another orbit perpendicular to the applied field, but with the electron going around in the opposite direction, as in Figure 5.3. The angular speed of the electron in the absence of the field is still given by Equation (5.4). When the field is present, the magnetic force acting on the electron is now outwards. Following the same argument as in the first case we obtain for the new angular speed ω' of the electron

Hence in both cases the electrons acquire an additional orbital angular momentum $m_e \omega_L r^2$ in the direction of the applied field. This gives them both an induced magnetic moment of magnitude

Figure 5.3. The effect of a magnetic field on the circular orbit of an electron whose orbital angular momentum is opposite to the field. The angular velocity ω_0 is decreased by the amount ω_L.


B
+Ze
r
-e
$\omega = \omega_0 + \omega_L$
-ev x B
v



from Equation (5.2). The induced moment is opposite to the increase in angular momentum, i.e. opposite to the applied field, and

(5.6)

on substituting for the Larmor angular frequency ωL.

An atom with atomic number Z has Z electrons. These electrons are in orbits with different radii and with different inclinations to the applied field. Taking the average of the contributions of all the electrons, it can be shown that the effective induced dipole moment per atom is given by

(5.7)

Where is the mean square radius of the electron orbits.

The magnetization of the diamagnetic material is now obtained from Equation (5.1).

(5.8)

A determination of the magnetization for a given field B inside a sample can thus provide an estimate of the radius of the atoms.

The above treatment of diamagnetism serves to convey its physical origin and also, in Equations (5.7) and (5.8), gives the same answers as a more correct approach. However, the discussion lacks rigour in two important respects. First we asserted in section 4.4.3 that the angular momentum of an electron in an orbit is quantized in units of $h/2\pi$, and this appears to contradict the conclusion that when a field B is applied the angular momentum changes by an amount depending on B. A proper quantum mechanical treatment of the angular momenta of charged bodies in the presence of electromagnetic fields removes this difficulty. The second problem with our simple picture is that even if we restrict ourselves to classical considerations, a rigorous classical treatment predicts zero diamagnetism, and indeed zero magnetization of any sort, for a system of particles in thermal equilibrium. We reemphasize that it requires quantum mechanics for a proper discussion of magnetism. For this the reader should consult more specialized texts*.

The field B in Equations (5.7) and (5.8) is strictly speaking the local magnetic field at the position of the atom, i.e. the total field less the contribution of the atom itself. However, for non-ferromagnetic materials the atoms have only a very small magnetic effect—for most practical purposes the local field can be taken to be the same as the macroscopic field B in the absence of the sample.

We have not explained the mechanism whereby the electrons in the orbits shown in Figures 5.2 and 5.3 acquired the extra angular momentum $m_e\omega Lr^2$. This is discussed in the next chapter.

5.1.2 Paramagnetism

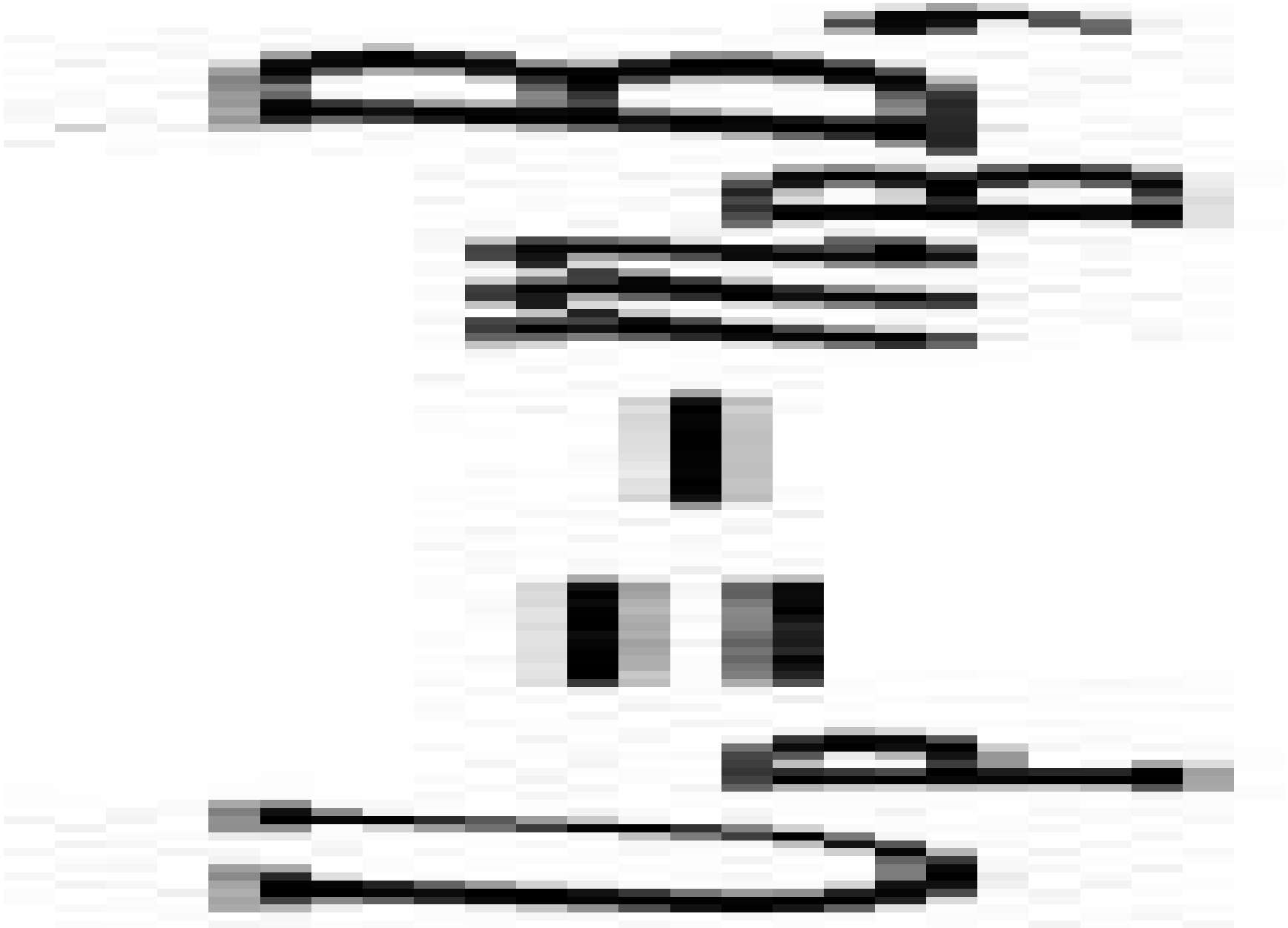
Some atoms and molecules have a non-zero electronic angular momentum, and thus have a permanent magnetic dipole moment. The orientation of the dipole

moments of individual molecules in a liquid or a gas is random in the absence of a magnetic field. When a field is applied there is a tendency for them to line up in the field direction and provide a paramagnetic effect. This process is analogous to the alignment of polar molecules in an electric field as discussed in section 2.2.2. The molecules will also acquire induced magnetic dipole moments, but this diamagnetic effect is usually smaller than the paramagnetism due to the permanent moments.

The derivation of the magnetization of a paramagnetic gas or liquid in terms of the permanent molecular magnetic dipole moments follows exactly the arguments presented in section 2.2.2. The theory given there was a good approximation because the potential energy of an electric dipole in an electric field was small compared with the thermal energy (~ kT, where k is Boltzmann's constant, and T the absolute temperature) of a molecule. We can easily show that the same is true for the potential energy of a microscopic magnetic dipole in a magnetic field. Permanent atomic magnetic dipole moments are of the same order of magnitude as the resultant orbital moment of the atomic electrons. This moment is given by Equation (5.3). The resultant orbital angular momentum is typically no more than a few times (≡ Planck's constant divided by 2π), and so a typical atomic dipole moment has a magnitude about equal to the quantity (e /2me). Atomic magnetic dipole moments are often expressed in units of this quantity, which is called the Bohr magneton, and given the symbol μB,

(5.9)

The numerical value of μB is 9.273×10^{-24} Am². The potential energy of a magnetic dipole moment mB equal to one Bohr magneton, aligned with a field of 10 T (a very large field) is given by Equation (4.26)



and is equal to ~ 10− 22 joules, or ~ 6 × 10− 4 eV. At room temperature the thermal energy kT is about 2.5 × 10− 2 eV, and so the potential energy of atomic magnetic dipoles in a magnetic field is almost always much less than kT.

The derivation of the magnetization can now be pursued in a manner identical to that followed in section 2.2.2. The result is similar to Equation (2.13),

(5.10)

In this formula m is the permanent dipole moment of the molecules and N is the number of molecules per unit volume. This result corresponds to an average dipole moment per molecule in the direction of the field

(5.11)

Again the field B should strictly speaking be the local field Blocal. However, as for diamagnetic materials, the two fields are nearly the same and for most practical purposes they need not be distinguished.

The total magnetization of a paramagnetic material must include the induced moments, and adding their contribution to Equation (5.10) we obtain

(5.12)

Let us substitute typical numbers into this equation in order to estimate the relative magnitudes of the paramagnetic and diamagnetic terms. For a material with Z equal to 50, and atoms which have a mean radius of 10− 10m and permanent moments of one Bohr magneton, the ratio of the diamagnetic to the paramagnetic term in Equation (5.12) is about one half at room temperature. For lighter atoms the ratio is correspondingly smaller, and it is a safe general rule that for liquids and gases made up of atoms with a permanent magnetic moment, the diamagnetism is weaker than the paramagnetism.

5.1.3 Ferromagnetism

Some substances, like iron, exhibit very large magnetic effects, and are called ferromagnetic. Below a certain temperature, called the Curie temperature, the spins of the conduction electrons in macroscopic sized portions called domains of certain conductors, are all aligned parallel. The cause of this is an interaction between the free electrons and those on the lattice ions. The total alignment within domains occurs only when the lattice ions have electrons in certain atomic orbitals. This happens e.g. in iron, cobalt and nickel and certain alloys of these metals. In a piece of unmagnetized iron the domains, which typically have volumes in the range 10− 10 to 10− 12 m^3, are oriented at random. In a weak external field, there is a net alignment of the domains in the field direction, giving a very large ferromagnetic effect. As the field increases the alignment of individual domains increases and those domains which are already aligned become enlarged at the expense of the others. These effects continue until saturation for the material is reached, when a significant fraction of all the conduction electron spins is lined up. This fraction depends on the structure of the material and on its previous physical treatment such as annealing or straining. Ferromagnetism also occurs in the rare-earth elements in which the electron interactions line up the dipole moments associated with lattice ions.

A local order, somewhat similar to that imposed on the conduction electrons or lattice ions in a ferromagnetic material, can also be found, for example in certain compounds containing ferromagnetic elements and in some nonferromagnetic

metals. This order results in adjacent dipole moments being equal in size and parallel but reversed in direction. These materials are called antiferromagnetic and have very small magnetism. Antiferromagnetic materials in which adjacent dipoles are of unequal size are called ferrimagnetics and have characteristics somewhat similar to ferromagnetics but more weakly manifested. Ferrites are typical ferrimagnetic materials and are complex compounds containing iron; they usually have high resistivity which makes them very useful in many applications.

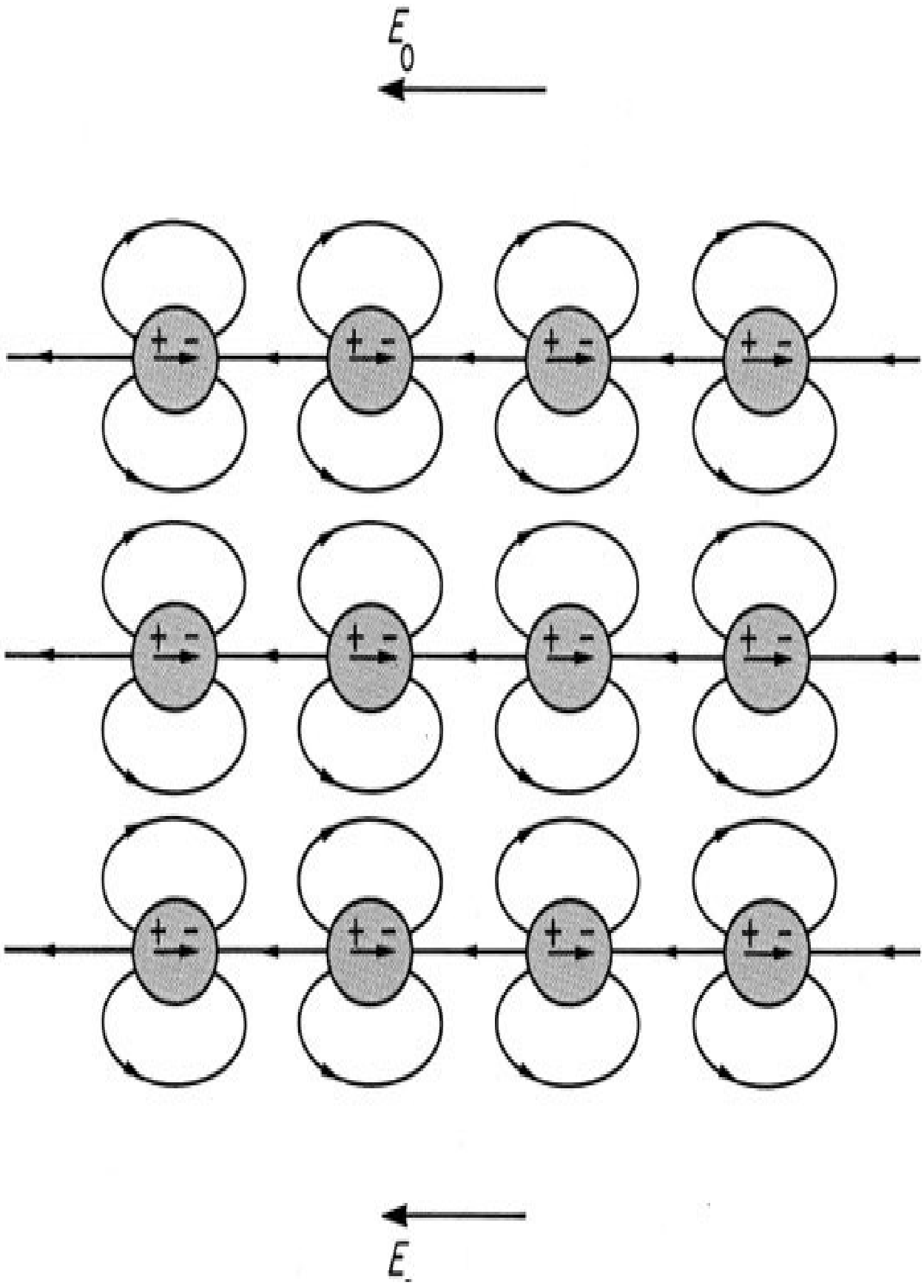
We will not discuss magnetism further in any detail at the atomic level. For a proper discussion we repeat the earlier message that the reader should consult more specialized texts. When discussing ferromagnetic materials in this book, we will from now on be concerned only with their macroscopic behaviour. This behaviour is given by the magnetization curve of a sample, which shows how the magnetization M varies with the field B in the sample. These curves are discussed later in section 5.4.1.

5.2 THE MACROSCOPIC MAGNETIC FIELD INSIDE MEDIA

If a substance is placed in a magnetic field arising from currents in nearby conductors, there is a field inside the material and the atoms or molecules are magnetized. The atomic magnetic field, Batomic, varies rapidly with position, even inside a single atom, and has contributions from both the atomic constituents of the material and from the nearby currents. The macroscopic field B in the medium is the average of the field Batomic taken over volumes very small on a macroscopic scale but containing very many atoms. Changes of field on the atomic scale are smoothed out and the macroscopic field is a relatively slowly and smoothly varying function of position. The magnetizing field acting on a molecule is the field Blocal. This is equal to the field Batomic less the contribution from the magnetization of the molecule itself, analogously to the field Elocal in electrostatics.

The field B is made up of two contributions. One is the original field B0 present when the specimen was absent. (This assumes that the current distributions which give rise to the field B0 are unaffected by the magnetization of the sample.) The other contribution is the field BM due to the magnetization M of the sample. The total field B is the sum of the fields from the two sources,

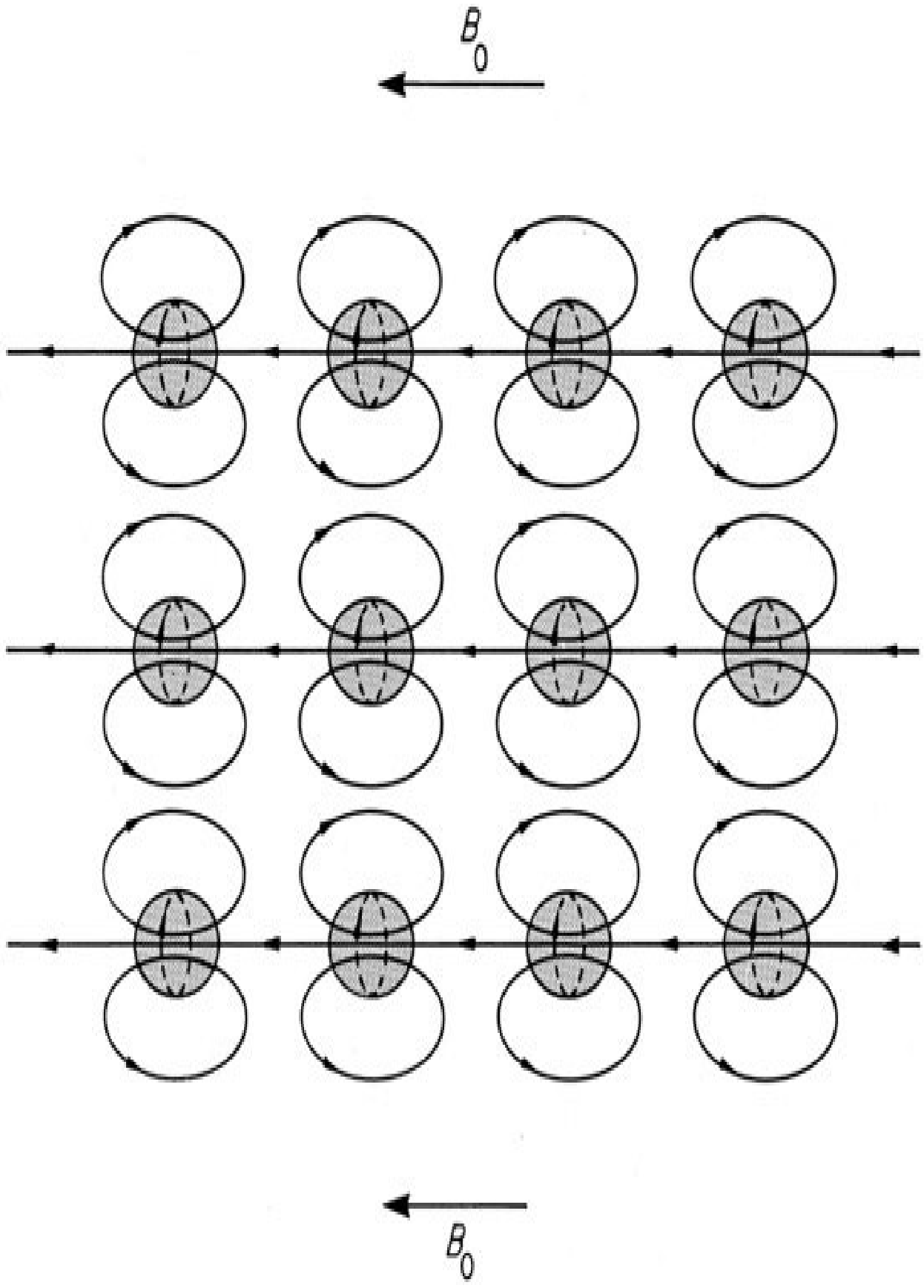
Figure 5.4. A schematic picture of polarized molecules within a dielectric material placed in an electric field. The electric field within the material is less than the original electric field E0 because the average field produced by the polarized molecules is in the opposite direction to E0.


E_0
E

(5.13)

Although the difference between the fields B and B0 is small for non-ferromagnetic materials, the sign of the difference is interesting. For paramagnetic materials it is found that the value of the macroscopic field inside is greater than the original field. For diamagnetic materials the value of the field B is less than that of the field B0. In the case of a dielectric in an electric field the polarization P is always in the same direction as the total electric field E, and the effect of the polarization is to reduce the initial field E0 present before the dielectric was placed in position. In the analogous case of a paramagnetic sample placed in a magnetic field, the magnetization is also in the direction of the total magnetic field, but acts to increase the field present originally. The difference between the two situations can be illustrated with simplified microscopic pictures of the interiors of the two specimens, keeping in mind the basic difference between the electrostatic field and the magnetic field. The lines of the fields Batomic and B are always continuous, whereas the lines of the electrostatic field begin or end on charges. Figure 5.4 shows schematically the atomic field due to atoms polarized within a dielectric. There is a reversal of the direction of the electric field within the atomic dipoles, and as we found in section 2.2, the net effect of the polarization is to make the macroscopic electric field E less than the field E0. Figure 5.5 shows the circulating current loops that represent the effective magnetic dipoles of the atoms of a magnetized paramagnetic material. In this case there are no changes of direction of the atomic magnetic field within the atoms. We shall show in the next section that the average atomic field in the magnetized material, i.e. the macroscopic field B, is greater than the field B0.

Figure 5.5. A schematic picture of magnetized molecules within a paramagnetic material placed in a magnetic field. The magnetic field within the material is greater than the field B0 originally present because in this case it turns out that the average field produced by the magnetized molecules is in the same direction as B0.


B_0
B_0

The field BM in ferromagnetic materials can be several orders of magnitude greater than those in other substances. The total field B inside a ferromagnetic sample is usually very much larger than the original field.

5.2.1 The surface currents on a uniformly magnetized body

In this section we discuss how a magnetized body can be replaced by currents in order to determine the magnetic field arising from the magnetization. This is a necessary step in the chain of reasoning which leads to equations satisfied by the magnetic field in the presence of magnetic materials.

Let us first estimate the field BM in a simple example before proceeding to the general case. Consider a long cylindrical solenoid, having N turns per unit length of thin wire carrying a current I. The field B0 inside the solenoid is uniform, in a direction parallel to the axis, and has magnitude

(5.14)

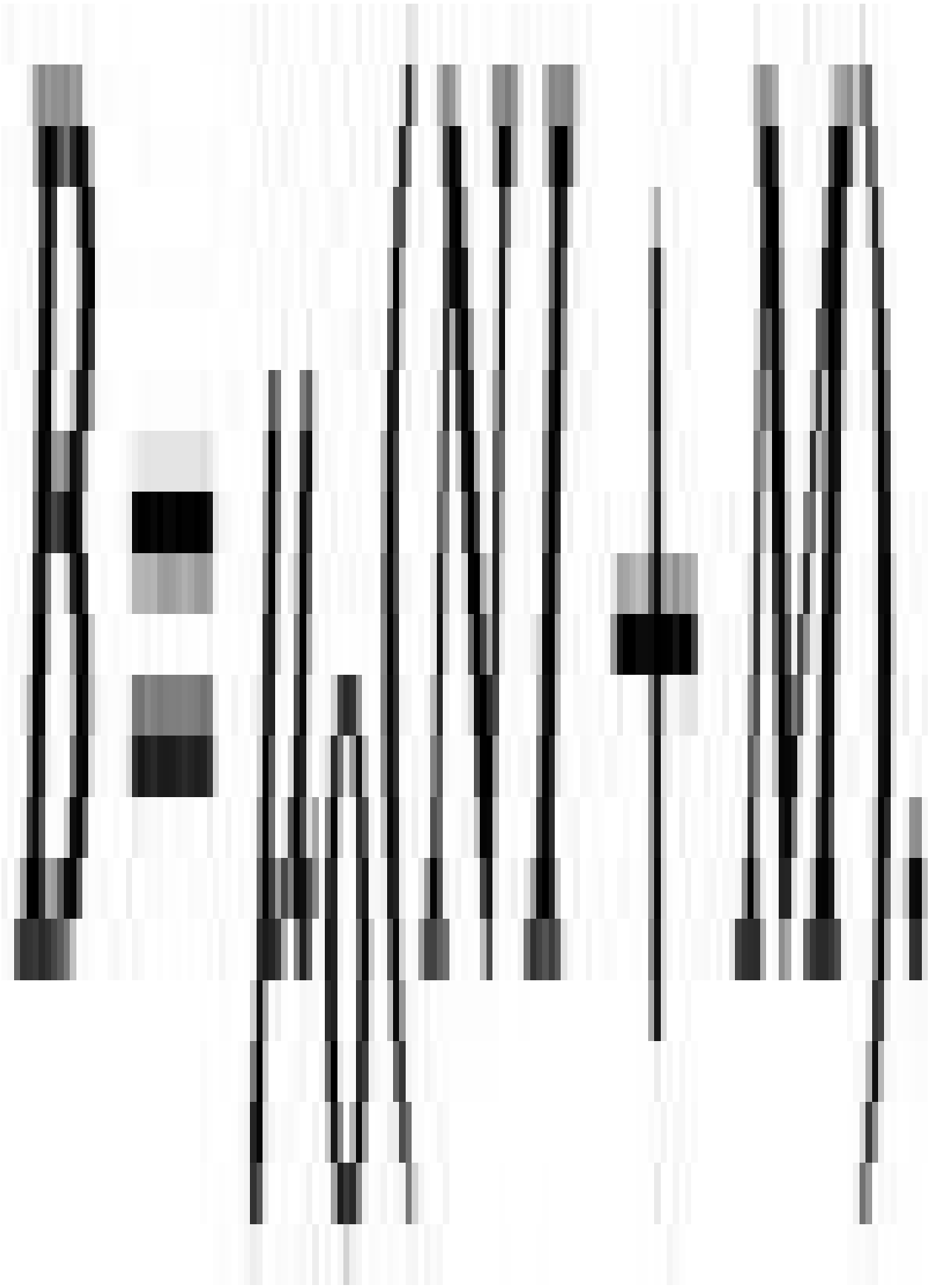
as given by Equation (4.36). If the solenoid is completely filled with a paramagnetic rod the magnetic field inside will remain uniform. The rod will acquire a uniform magnetization M in a direction along its axis. If the rod is divided into very thin slices of thickness δx perpendicular to the axis, each slice is equivalent to a uniform sheet of magnetic dipoles. The total magnetic dipole moment of each slice is $\pi r^2 \delta x M$, where r is the radius of the solenoid. As discussed in section 4.4.1, each individual dipole of the sheet can be replaced by a small current circulating in the plane of the disc of side δx. All the currents inside the disc cancel each other, leaving a current flowing around the disc's rim in a vanishingly thin layer on a macroscopic scale. This current is called a

magnetization surface current. Let it have magnitude is per unit length of surface. The dipole moment $\pi r^2 is\delta x$ due to the magnetization current around the rim of the disc of side δx is equal to the dipole moment $\pi r^2 \delta x M$ of the whole disc, hence

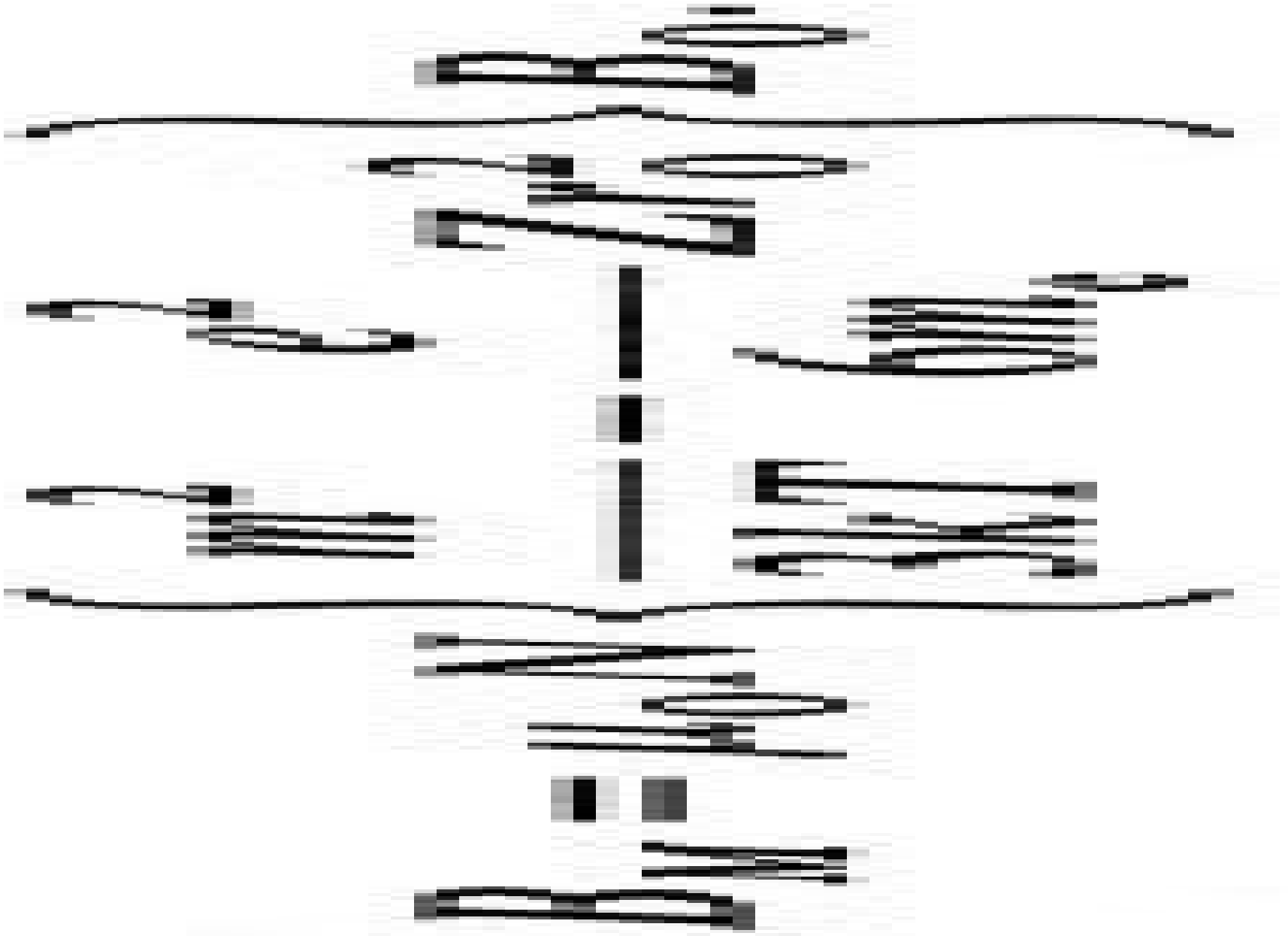
A unit length of the magnetized rod is thus equivalent (for the purpose of calculating the magnetic field to which it gives rise) to a current flowing around the outside of the rod of magnitude per unit length equal to the magnitude of the magnetization. Using the same arguments as in section 4.4.3 to derive the field due to the current in the solenoid coils, it can be seen that the magnetization surface current gives rise to a uniform field inside the solenoid of magnitude

(5.15)

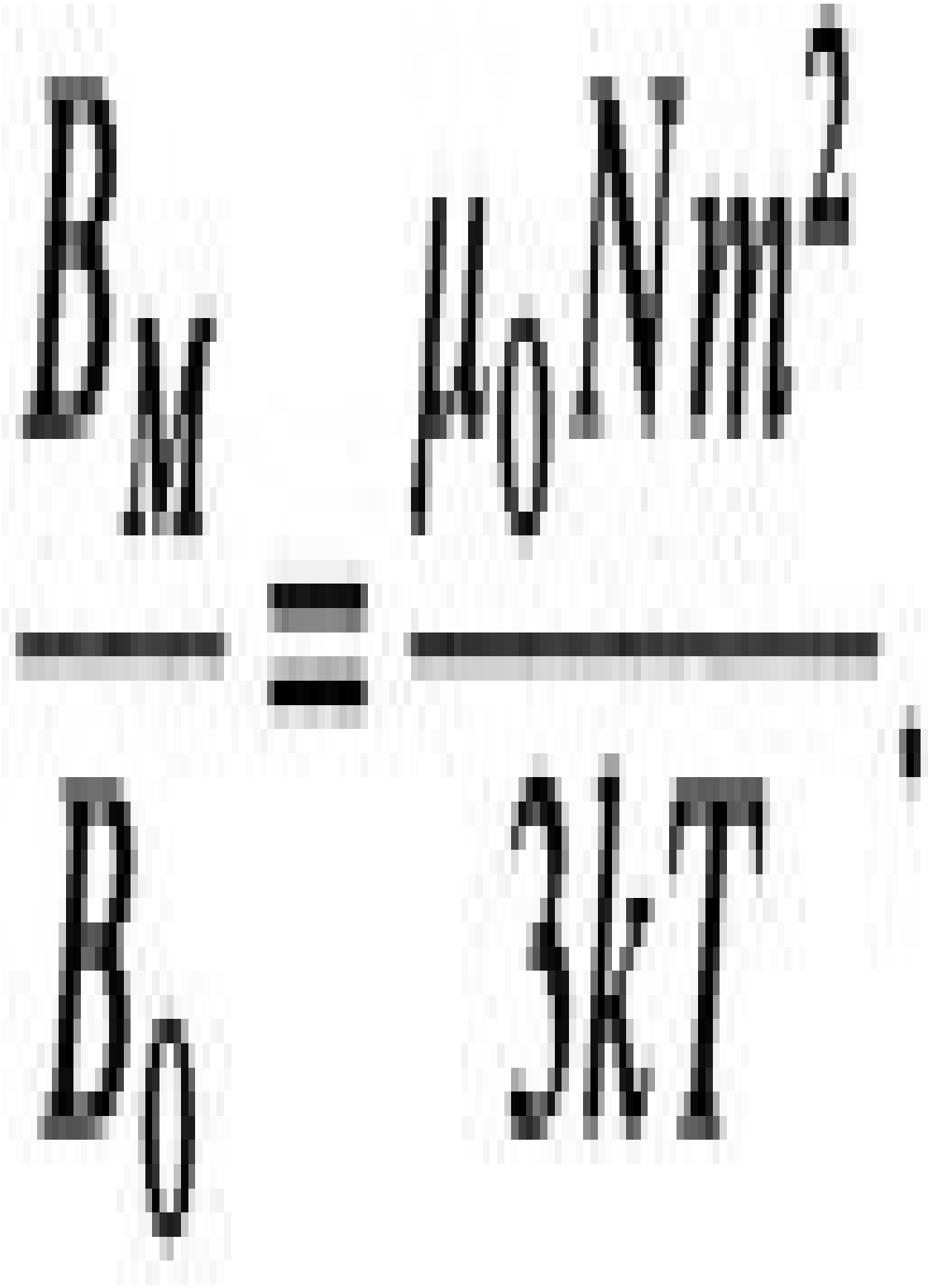
The total field B inside the solenoid is the sum of this field and the field B0 given by Equation (5.14),



We can use the results obtained above to show that the local field in paramagnetic or diamagnetic materials is usually very nearly the same as the macroscopic field. To do this we have to show that the magnetization of such materials produces only a very small field BM compared with the original field B0 in which the material was placed. If this is so Equation (5.13) becomes B B0, and the local field to the same approximation is also equal to the original field. In the simple example considered above let us first assume that the field BM produced by the magnetized rod is indeed very much smaller than the field B0. In this case the magnetization can be obtained from Equation (5.12) by replacing the field B by B0, and the field arising from the magnetization is given by Equations (5.12) and (5.15) to be



The diamagnetic term in this equation is usually less than the paramagnetic term and we will neglect it. With this approximation



Taking the permanent magnetic dipole moment m equal to one Bohr magneton, and the number of molecules per unit volume N equal to 5×10^{28} m−3 (a typical value for condensed matter), the ratio BM/B0 is given by the above equation to be $3 \times 10-4$ at room temperature. If the diamagnetic term had been included the ratio would have been even smaller. Our conclusions are thus consistent with the original assumption, and we have shown that to a good approximation the field BM can be neglected compared with the field B0. The magnetization of nonferromagnetic materials makes very little difference to the magnetic field.

5.2.2 The distributed currents within a magnetized body

We will now deal with the more general case. The magnetization current flowing around the surface of a uniformly magnetized material is analogous to the polarization charges on the surface of a uniformly polarized dielectric. A non-uniformly polarized dielectric may give rise to a volume distribution of charge, in addition to charges over its surface. Similarly a non-uniformly magnetized material may give rise to magnetic fields as though it carried a volume distribution of current, in addition to currents over its surface. In order to see how such volume distributions of current arise, consider the slab of magnetized material shown in Figure 5.6. The magnetization is everywhere in the y-direction. Its magnitude is independent of the coordinates y and z, but increases as the coordinate x increases. The effective microscopic moments that constitute the magnetization may again be regarded as circulating atomic currents. A suitably sized current circulating in a loop of atomic radius will give the same magnetic effects on a macroscopic scale as the average effective magnetic dipole moment of the atoms. Figure 5.7 shows the atomic current loops over part of the plane ABCD cut through the magnetized slab. The circulating currents in the loops increase as the coordinate x increases, corresponding to the increasing effective dipole moments of the atoms. The result is a net upward atomic current along any line parallel to the z-axis, since the current down just to the left of the line is less than the current up just to the right. Averaging over volumes which are vanishingly small on a macroscopic scale but which contain very many atoms gives a macroscopic picture in which upward currents persist throughout

the slab, increasing towards the right. These macroscopic currents are called magnetization currents. They are specified by the magnetization current density jM and give rise to magnetic fields just like the currents due to motion of free charges discussed in Chapter 4.

Figure 5.6. A volume within a magnetized magnetic material. The magnetization is in the y-direction and varies only with the coordinate x, increasing as x increases.

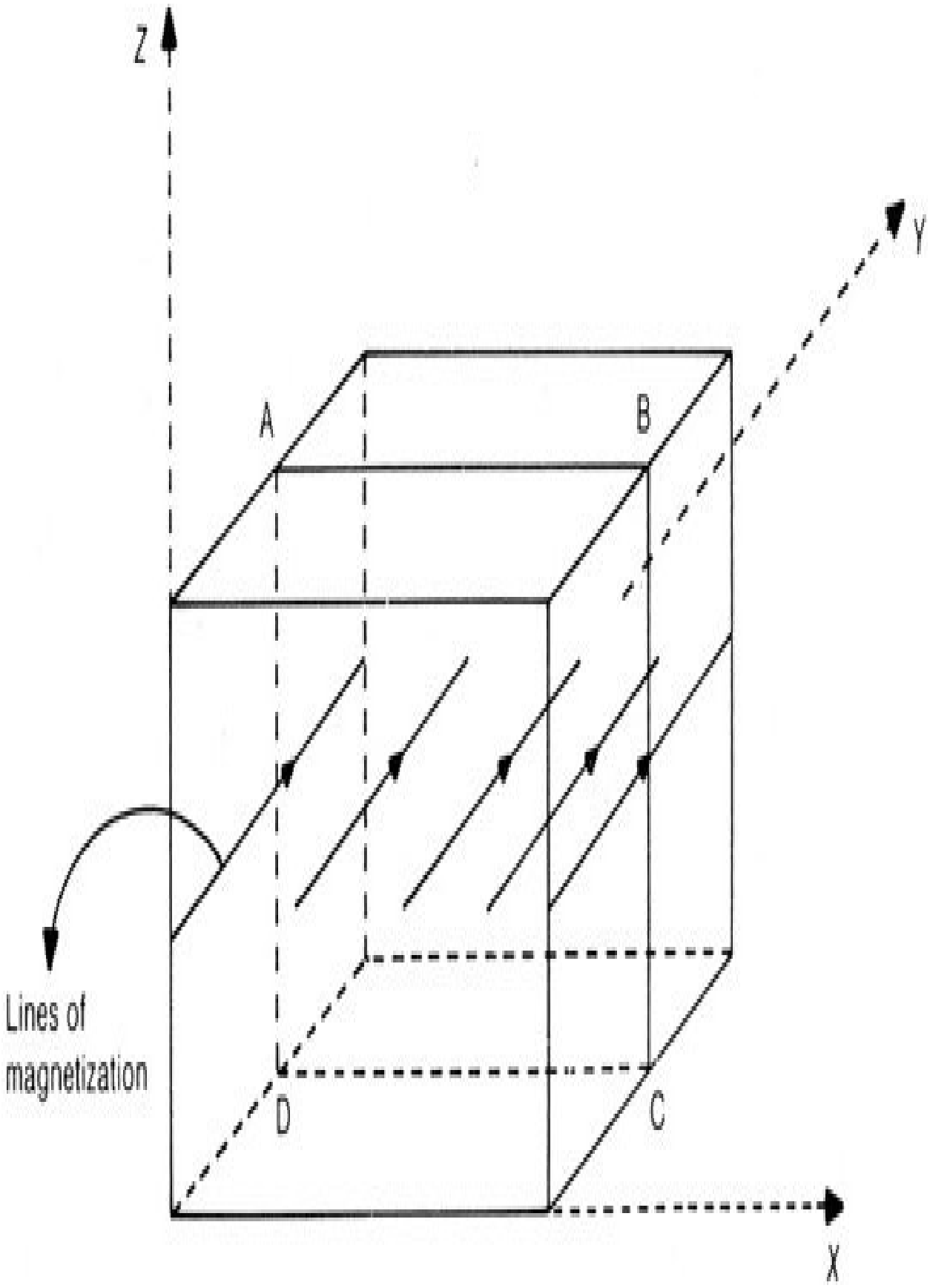
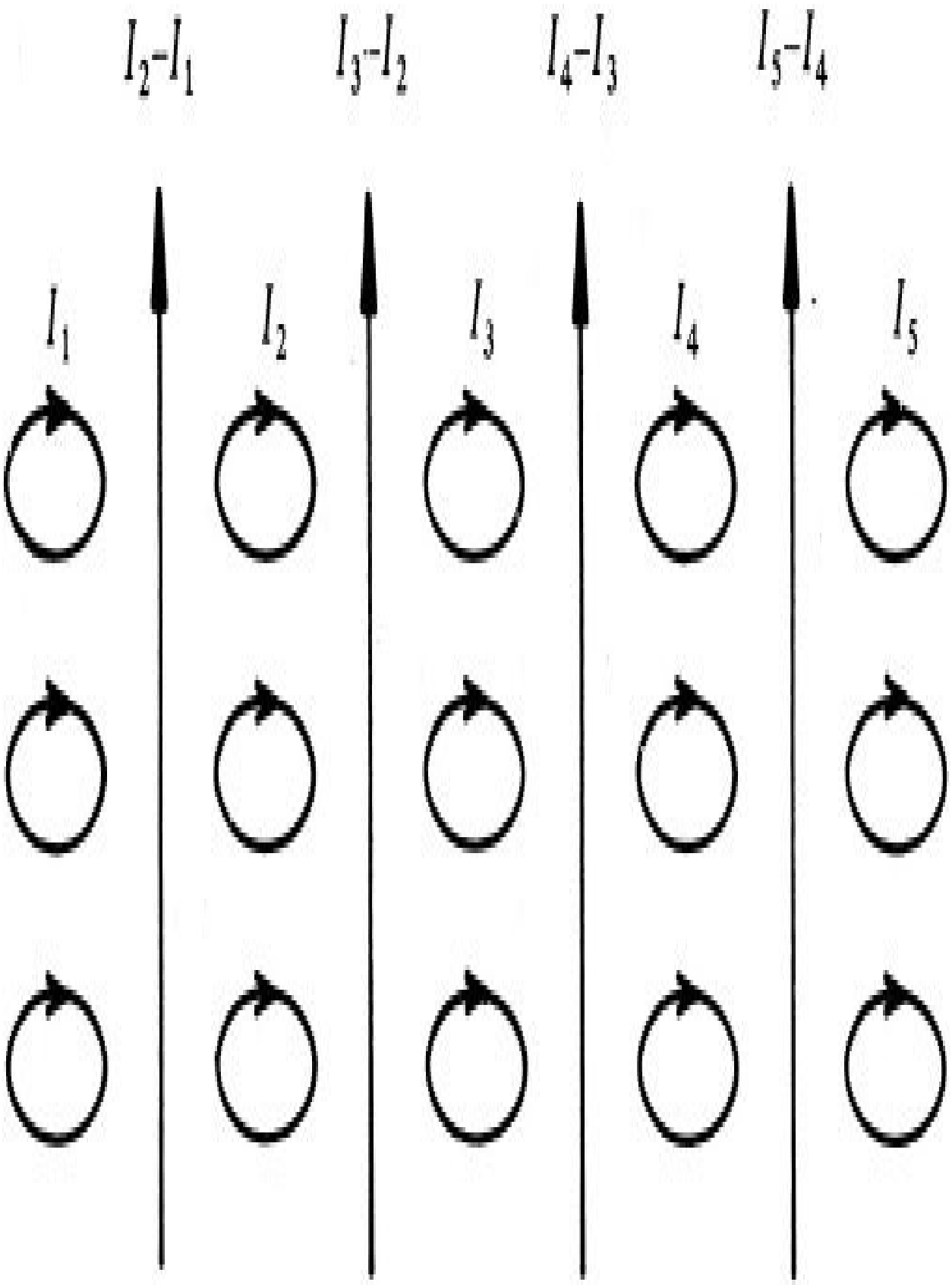
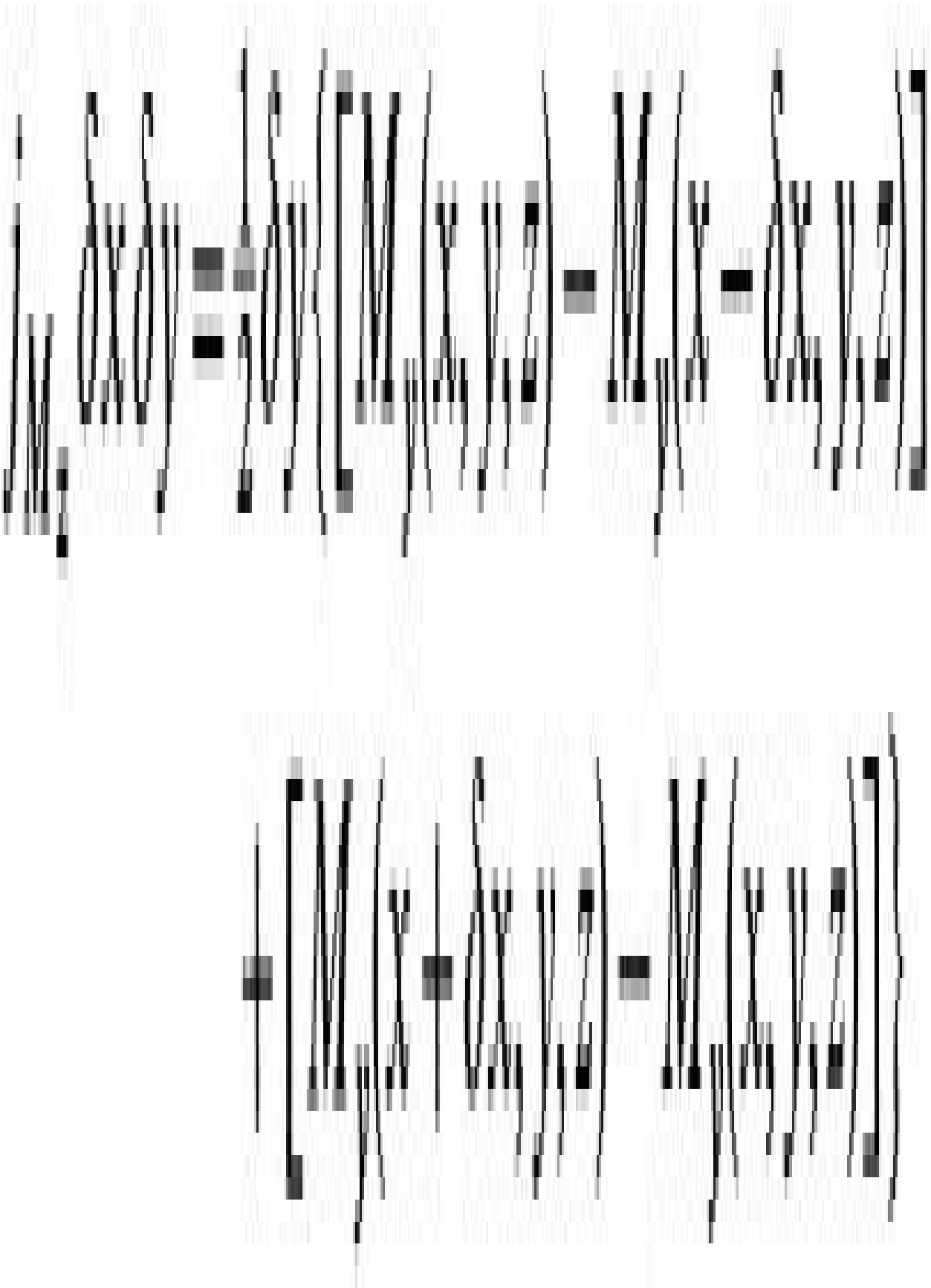

Z
Y
A
B
Lines of
magnetization
D
C
X

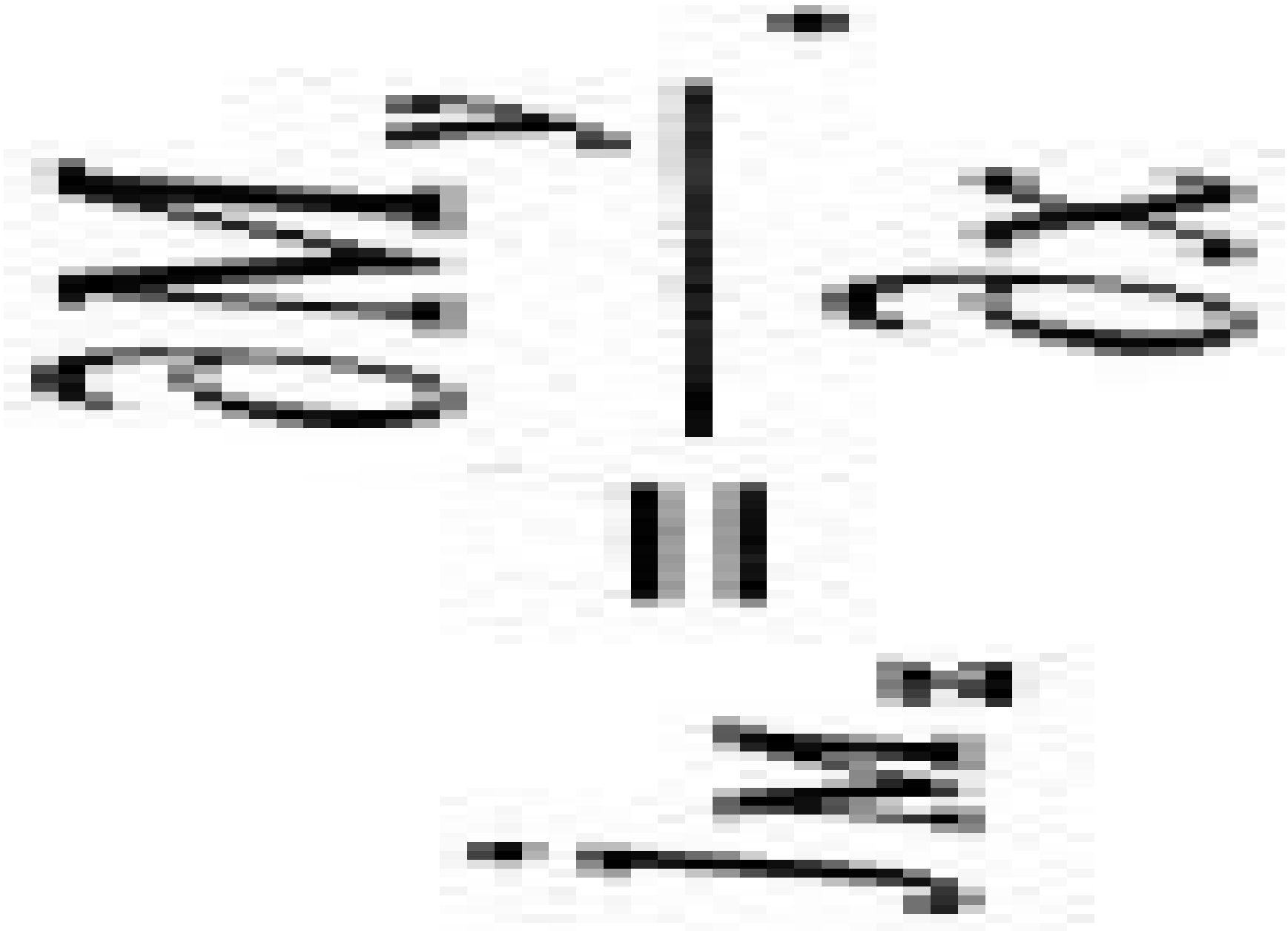
Figure 5.7. A schematic representation of the atomic current loops which give rise to the magnetization over part of the plane ABCD of the magnetized slab shown in Figure 5.6. The currents increase from left to right.


I_2-I_1
I_3-I_2
I_4-I_3
I_5-I_4
I_1
I_2
I_3
I_4
I_5

Let us continue to examine the magnetized sample shown in Figure 5.6. The magnetization current density is everywhere in the z-direction with size , and thus the current flowing upward over the element A centred on the point (x, y, z) shown in Figure 5.8 is equal to $\delta x \delta y$. This current can be expressed in terms of the magnetization distribution within the sample. The magnetic dipole moment of the element A is $M_y(x,y,z)\delta\tau$ where $\delta\tau$ is its volume, equal to $\delta x \delta y \delta z$, and M_y the magnetization. This dipole moment is equivalent to a current of size $M_y \delta y$ flowing around the loop of area $\delta x \delta z$ shown in Figure 5.8. We obtain the net upward current over element A by considering two terms. The first term is the difference between the current $M_y \delta y$ up over the left-hand face of element A and the current down $M_y(x - \delta x, y, z)\, \delta y$ over that face due to the magnetization of the cube to the left centred on the point $(x - \delta x, y, z)$. The second term is the difference between the current $M_y (x + \delta x, y, z)\, \delta y$ up over the right-hand face of element A due to the magnetization of the element to the right centred on $(x + \delta x, y, z)$ and the current down over that face from the magnetization of element A itself. The sum of these two terms is shared between two elements, hence the net upward current over element A is given by

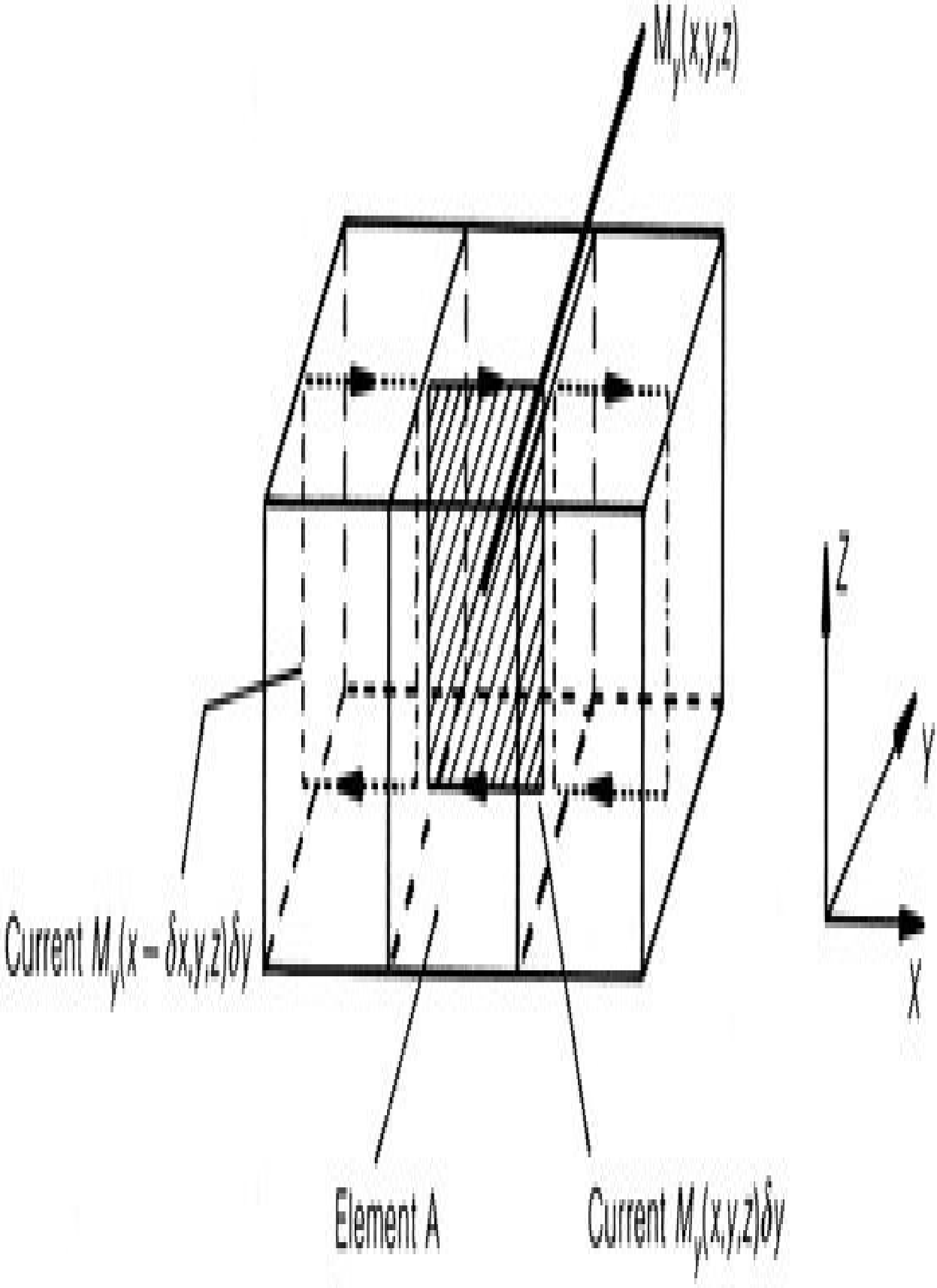


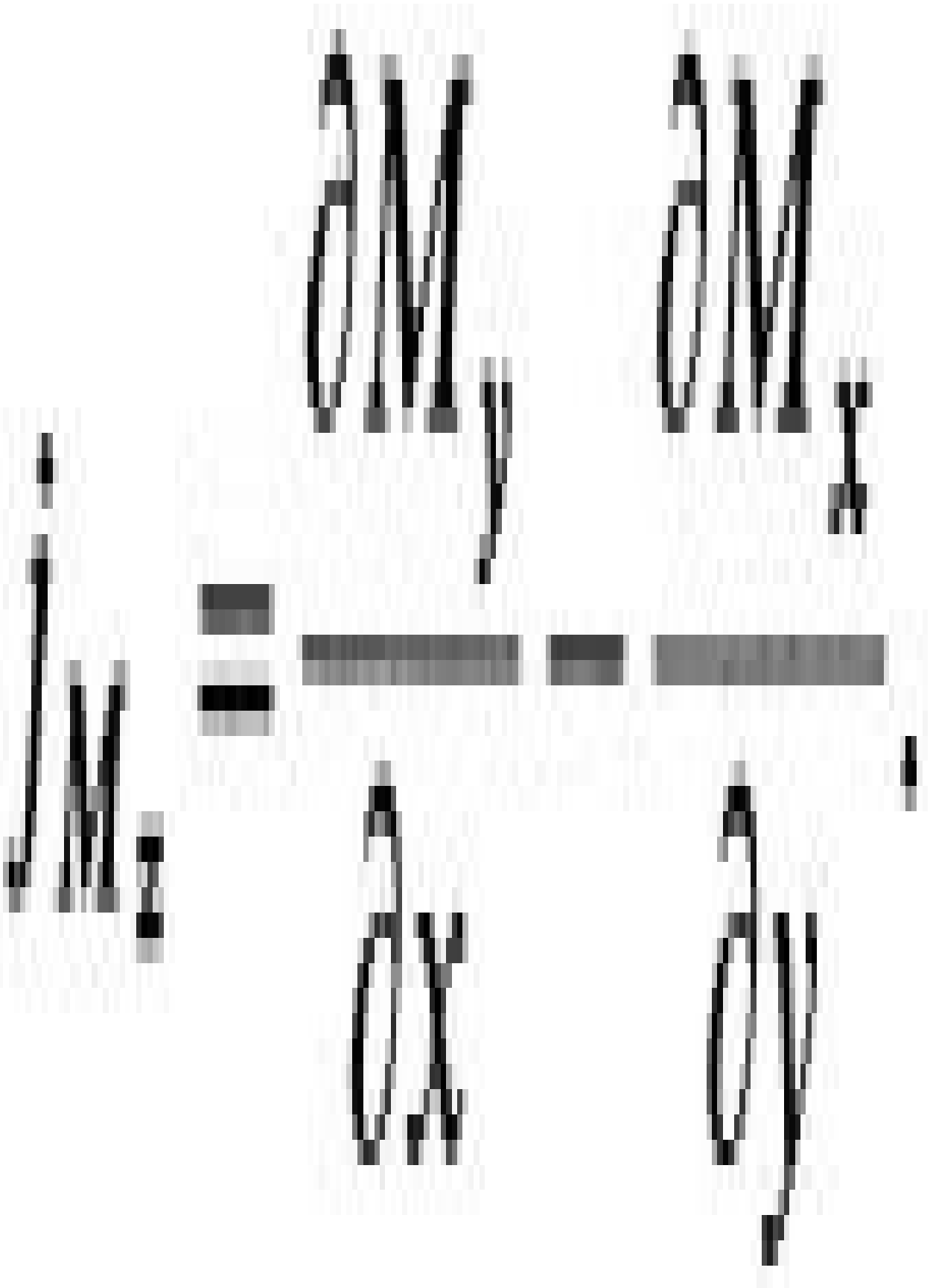
from which we obtain



In the more general situation where the magnetization M varies in a complicated way the z-component of the magnetization current density can be obtained by extending the above argument. We must now include contributions from both the x- and y-components of the magnetizations of element A and adjacent elements. There is no contribution to from the z-components since these correspond to current loops in the x–y plane. The x-component of the magnetic dipole moment of element A corresponds to a current of size $Mx(x,y,z)\delta x$ around the loop of area $\delta y \delta z$ over the surface of the element. This current and similar currents in the two adjacent elements at the points $(x,y - \delta y,z)$and $(x,y + \delta y,z)$ contribute to the net upward current over element A, and by an analysis similar to that above it can be shown that they contribute an amount $\partial Mx/\partial y$ to the z-component of the magnetization current density. Adding this to the amount found earlier we obtain the general expression

Figure 5.8. Very small macroscopic volume elements within a magnetized material. The element A is centred on the point (x, y, z).


$M_y(x,y,z)$
Z
Y
X
Current $M_y(x-\delta x,y,z)\delta y$
Element A
Current $M_y(x,y,z)\delta y$



The right-hand side of this equation is the z-component of the curl of the magnetization M. (See Equation (4.44).) So we may write

(5.16)

The x- and y-components of the vector curl M can be shown in a similar fashion to be equal to the x- and y-components of the magnetization current density jM:

(5.17)

(5.18)

Equations (5.16), (5.17) and (5.18) can be summarized in the form

(5.19)

This equation is the one we have been seeking in this section and tells us how to replace a volume distribution of magnetization by an equivalent current distribution throughout the volume.

Over the surface of a magnetized body there may be surface currents as described in the last section. These currents also give rise to magnetic fields. In the last section we saw that for a rod uniformly magnetized in a direction along

its axis the surface current per unit length was perpendicular to the magnetization in the direction given by the right-hand rule. The total surface current over a length l measured along the rod parallel to the magnetization was Ml. If we now consider a different situation where the magnetization is perpendicular to the surface the magnetization arises from currents flowing in loops parallel to the surface. Around the perimeter of any elemental surface element these currents cancel each other out and the net surface current at any point is zero. We thus see that when the vector M is perpendicular to the surface the surface current is zero, and when it is parallel to the surface the surface current is large. In general the surface current per unit length is a vector is with size and direction depending on the size and direction of the magnetization at the surface. The surface current per unit length is perpendicular to the vector M and the above observations suggest that it is proportional to the vector product of the magnetization M and the unit vector n in the direction of the outward normal to the surface, i.e.

(5.20)

In the discussion of the uniformly magnetized rod it was found that

(5.21)

hence Equation (5.20) becomes

(5.22)

Equations (5.19) and (5.22) give us complete prescriptions for replacing a magnetized body by equivalent currents which give the same magnetic field.

They are the analogues of Equations (2.17) and (2.3) which give the corresponding prescriptions relating the polarization charges to the dipole moment per unit volume within a dielectric material.

5.2.3 Magnetic susceptibility and atomic structure

We have established that the magnetization of a non-ferromagnetic material is proportional to the magnetic field B for isotropic, homogeneous materials.

We may thus write

(5.23)

where χB ls a dimensionless constant, called the magnetic susceptibility of the material.* Its value may vary with temperature, but at constant temperature the relationship between the magnetization and the field is linear.

For ferromagnetic materials the dependence of the magnetization on the field is non-linear. Also for most ferromagnetic materials there is no unique value of the magnetization corresponding to a given value of the field B. If the external field in which a piece of iron is situated changes, the resulting magnetization of the iron depends on the state of magnetization before the change. This phenomenon is called hysteresis and is discussed in section 5.4. For the remainder of this section we discuss paramagnetic and diamagnetic materials only.

One way in principle of measuring magnetic susceptibilities is to determine the

field near a sample magnetized by an external field. However, this method is not sensitive enough, since the change in the field arising from the presence of the sample is so small. Magnetic susceptibilities are usually measured by determining the force on a magnetized sample in an inhomogeneous magnetic field. This is discussed in the next chapter. Measurements of the magnetic susceptibilities of materials give information on their microscopic structures. Combining Equations (5.12) and (5.23) we obain the relation

(5.24)

for the susceptibility of a liquid or gas. The two contributions to the susceptibility can be separated by performing experiments at different temperatures, as discussed in section 2.2.2 for dielectric materials. Interpretation of the results gives the permanent moment m of the molecules, and the mean square radius of their electron orbits.

Table 5.1 gives the magnetic susceptibilities of several materials at a temperature of 293 K. Consider the noble gases as an example. Their atoms have no permanent magnetic dipole moments, since the total angular momentum of the electrons is zero for all the noble gas atoms. They all have the same number of atoms per unit volume at normal temperature and pressure, equal to 2.69×10^{25} m−3, and Equation (5.24) can be used to calculate their root mean square radii. Table 5.2 shows the calculated values. The sizes of the atoms are roughly equal as predicted by quantum mechanics. The sizes are grouped around a radius of about 0.55 Å. This is a surprisingly good estimate of the size of a noble gas atom in view of the simplicity of theory we have used.

As another example of the microscopic information which can be obtained from susceptibilities let us estimate the magnetic dipole moment of the oxygen molecule. At 293 K and atmospheric pressure oxygen is strongly paramagnetic. This behaviour can be contrasted with the weak diamagnetism of nitrogen. We will thus neglect the induced magnetism and attribute the susceptibility of

oxygen as given in Table 5.1 to the permanent dipole moments alone. With this approximation Equation (5.24) becomes

(5.25)

Table 5.1. The magnetic susceptibilities of some common substances at 293 K.

MATERIAL	SUSCEPTIBILITY χB
Gold	−3.6 × l0− 5
Copper	−0.98 × 10− 5
Germanium	−1.5 × l0− 5
Tungsten	+6.8 × l0− 5
Aluminium	+2.3 × 10− 5
Magnesium	+1.2 × 10− 5
Glass	−1.1 × 10− 4
Fused Quartz	− 6.2 × 10− 5
Nitrogen (76 cm Hg pressure)	− 6.7 × 10− 9
Oxygen (76 cm Hg pressure)	+1920.0 × 10− 9
Helium (76 cm Hg pressure)	−1.05 × 10− 9
Neon (76 cm Hg pressure)	−3.76 × 10− 9
Argon (76 cm Hg pressure)	−11.0 × 10− 9
Krypton (76 cm Hg pressure)	−11.1 × 10− 9
Xenon (76 cm Hg pressure)	−24.6 × 10− 9
Water	− 9.1 × 10− 6
Sodium Chloride	−1.38 × 10− 5

Table 5.2. The root mean square radii of the noble gas atoms obtained from their magnetic susceptibilities at normal temperature and pressure.

	Z	$\chi_B \times 10^{-9}$	$r_0 \times 10^{-10}$ m
He	2	− 1.0	0.575
Ne	10	− 3.8	0.486
Ar	18	− 11.0	0.620
Kr	36	− 16.1	0.530
Xe	54	− 24.6	0.536

Substitution of the appropriate numbers into this equation gives a value for the permanent moment m equal to ~2.6 × 10− 23 Am^2; i.e. about two and a half times a Bohr magneton.

5.3 THE FIELD VECTOR H

Suppose we need to calculate very precisely the magnetic induction field near the end of a short solenoid which has a rod of magnetic material inside it. How do we do such a calculation? The field is the addition of two parts, one part B0 due to the current in the solenoid, the other BM due to the magnetization of the rod. The part B0 can be calculated from the Biot–Savart law, but the contribution BM depends on the magnetization M of the material of the rod, and we do not know the distribution of magnetization, although we may know a relation between the magnetization M and the total magnetic field B at each point in the rod. Thus M depends on B, which is the quantity we want to calculate. The problem is most easily treated by introducing a third magnetic field vector, given the symbol H, and called the magnetic intensity or more usually the magnetic field H. The field vector H obeys simple boundary conditions and has the important property that its line integral around any closed path depends only on free currents caused by the flow of electrons in wires, for example, and is independent of magnetization currents. The use of the field H is analogous to the use of the electric displacement in electrostatics. There the field vector D was introduced in order to avoid direct reference to polarization charges. In electrostatics we worked with the vectors E and D, instead of E and the polarization P. Similarly with magnetic media it turns out to be easier to work with the vectors B and H than to work with the vectors B and M.

5.3.1 Ampère’s law for the field H

The line integral of the field B around a closed path is equal to the current I through the area defined by the path, times the permeability of free space μ0, as discussed in section 4.4.1. If the path encloses magnetized material there will in general be a magnetization current through the area defined by the path. This magnetization current gives rise to a magnetic field in the same way as currents due to the motion of free charges. Hence it must be included in the relationship involving the line integral of the field B.

We first consider a simple example before proceeding to the general case. Figure 5.9 shows part of a long solenoid filled with a rod of magnetic material. We will work out the line integral of the field B over the path ABCD, assuming as in section 4.4.3 that the field inside the solenoid is uniform. The line integral is

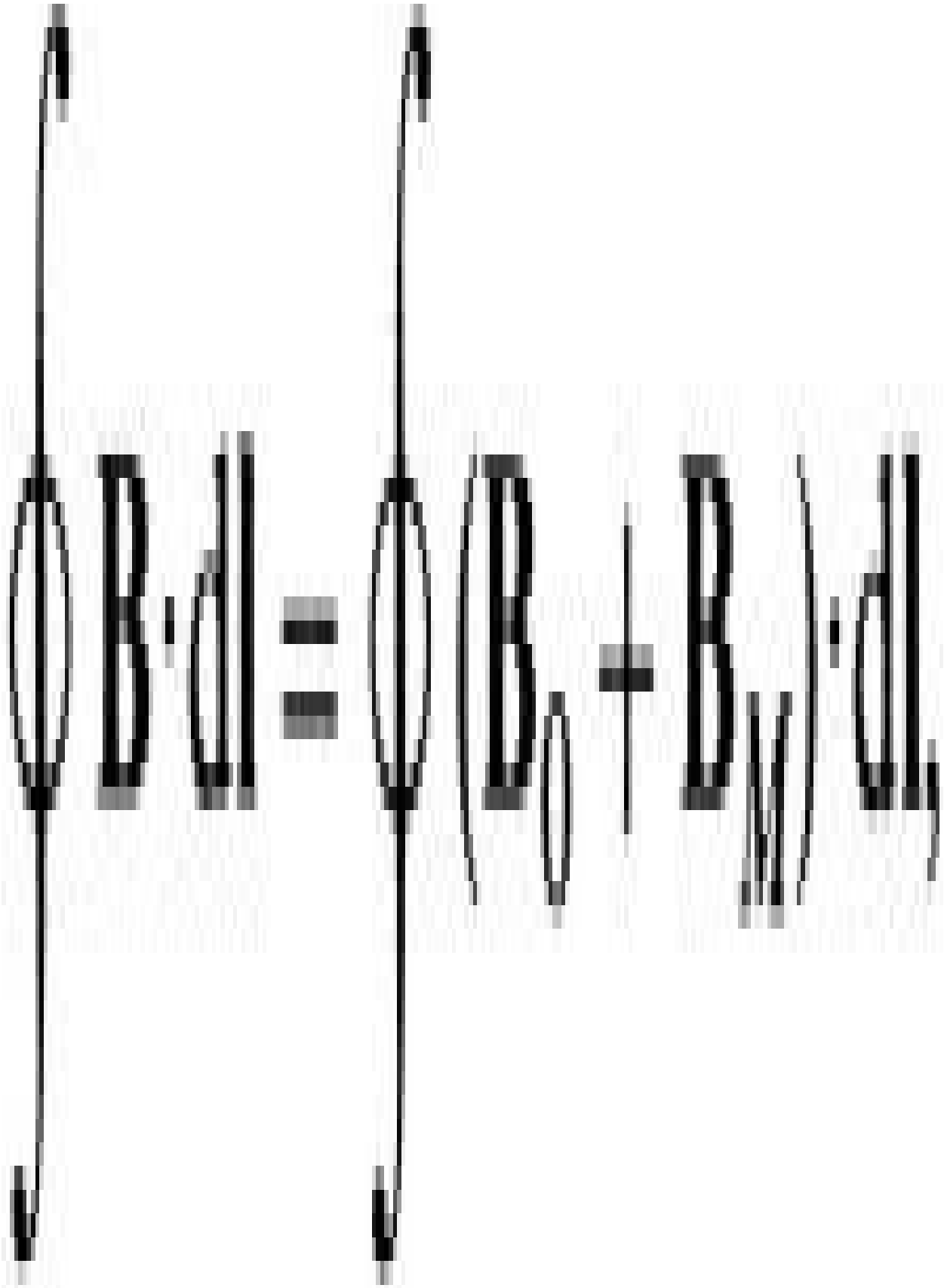
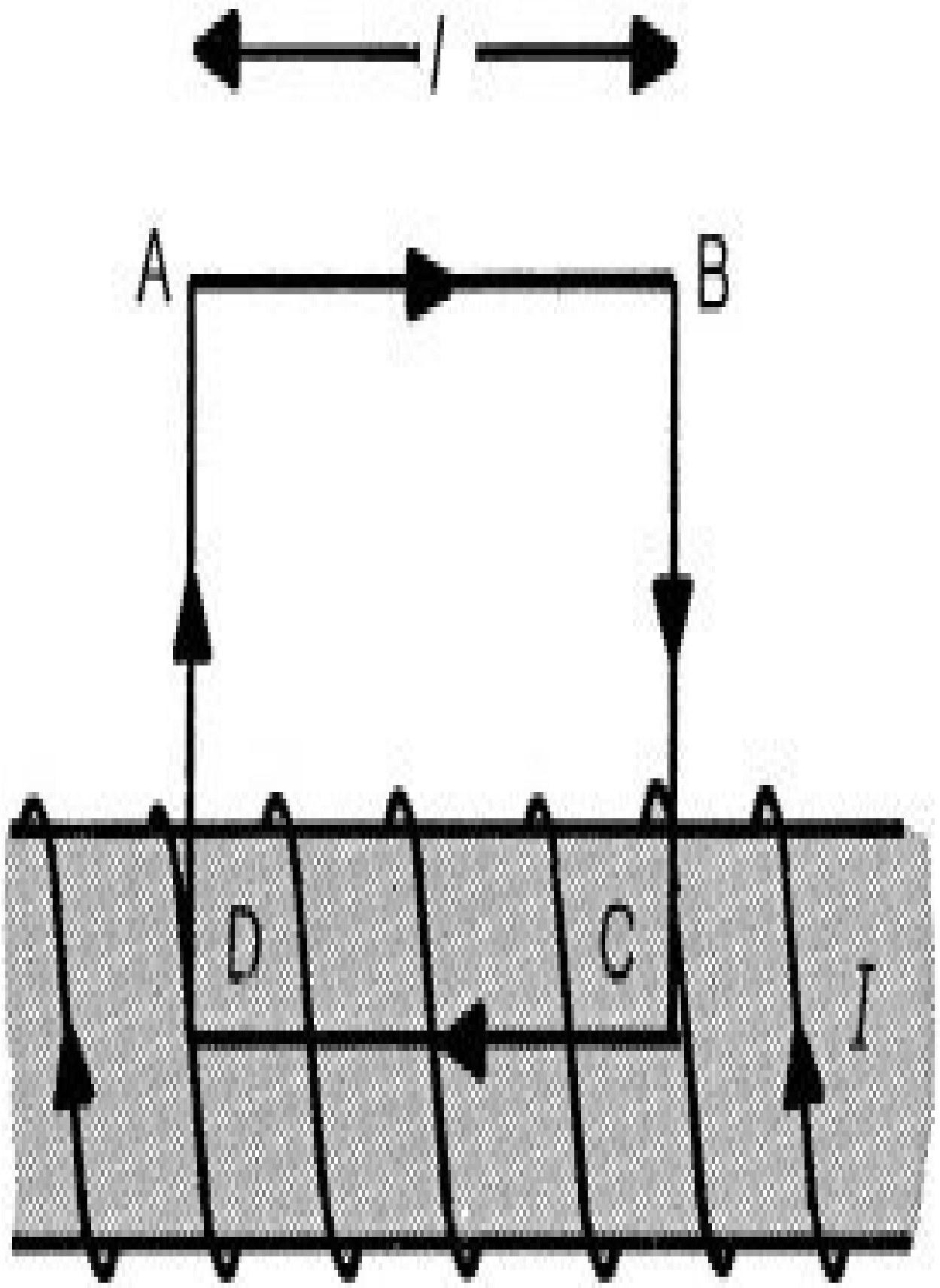
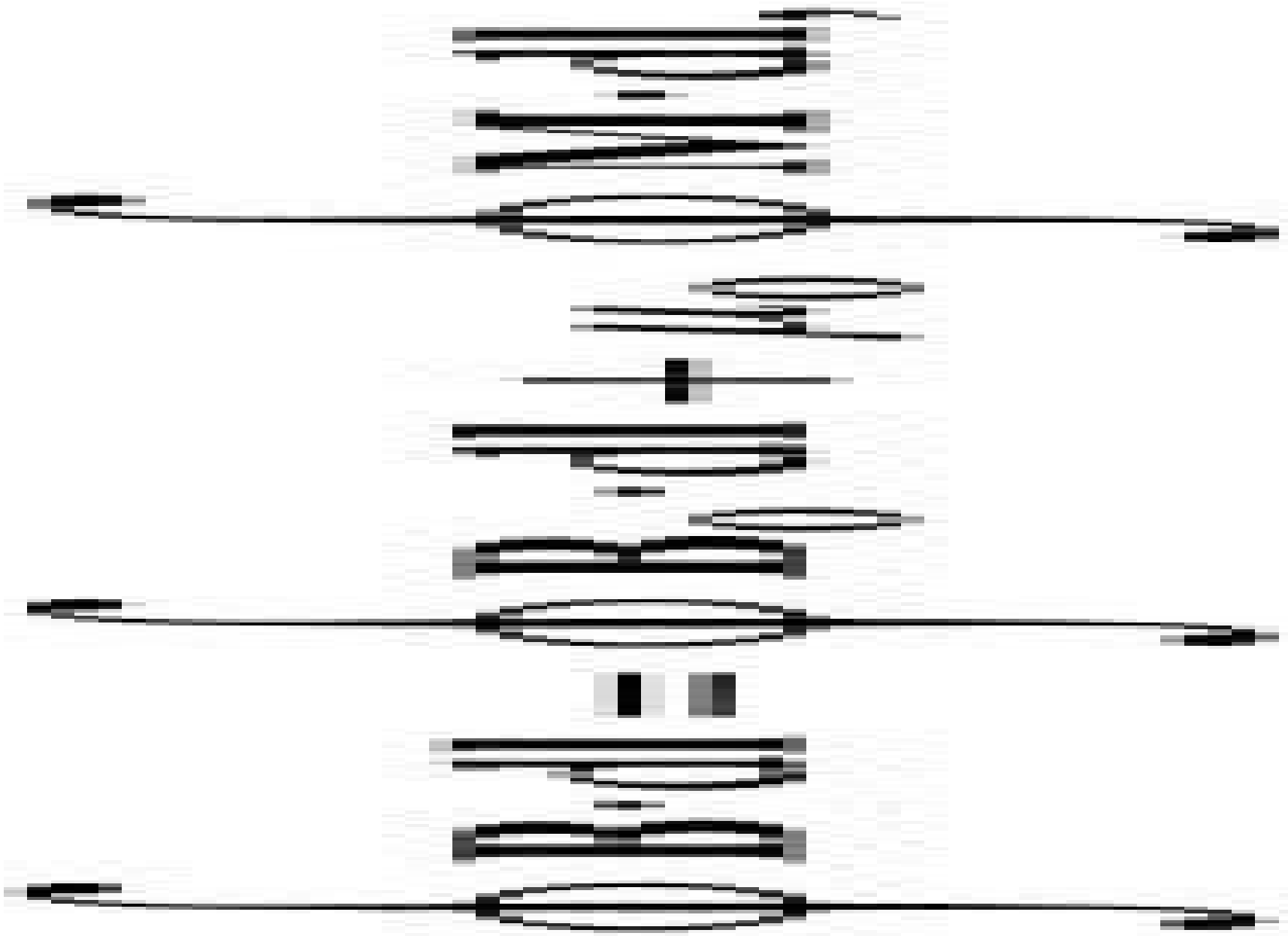


Figure 5.9. Part of a long solenoid completely filled with a rod of magnetic material. The rod is uniformly magnetized. The magnetic field within the rod is everywhere parallel to the axis, and is zero outside. The line integral of the field B over the closed loop ABCD has a contribution only from the part CD, of length l.

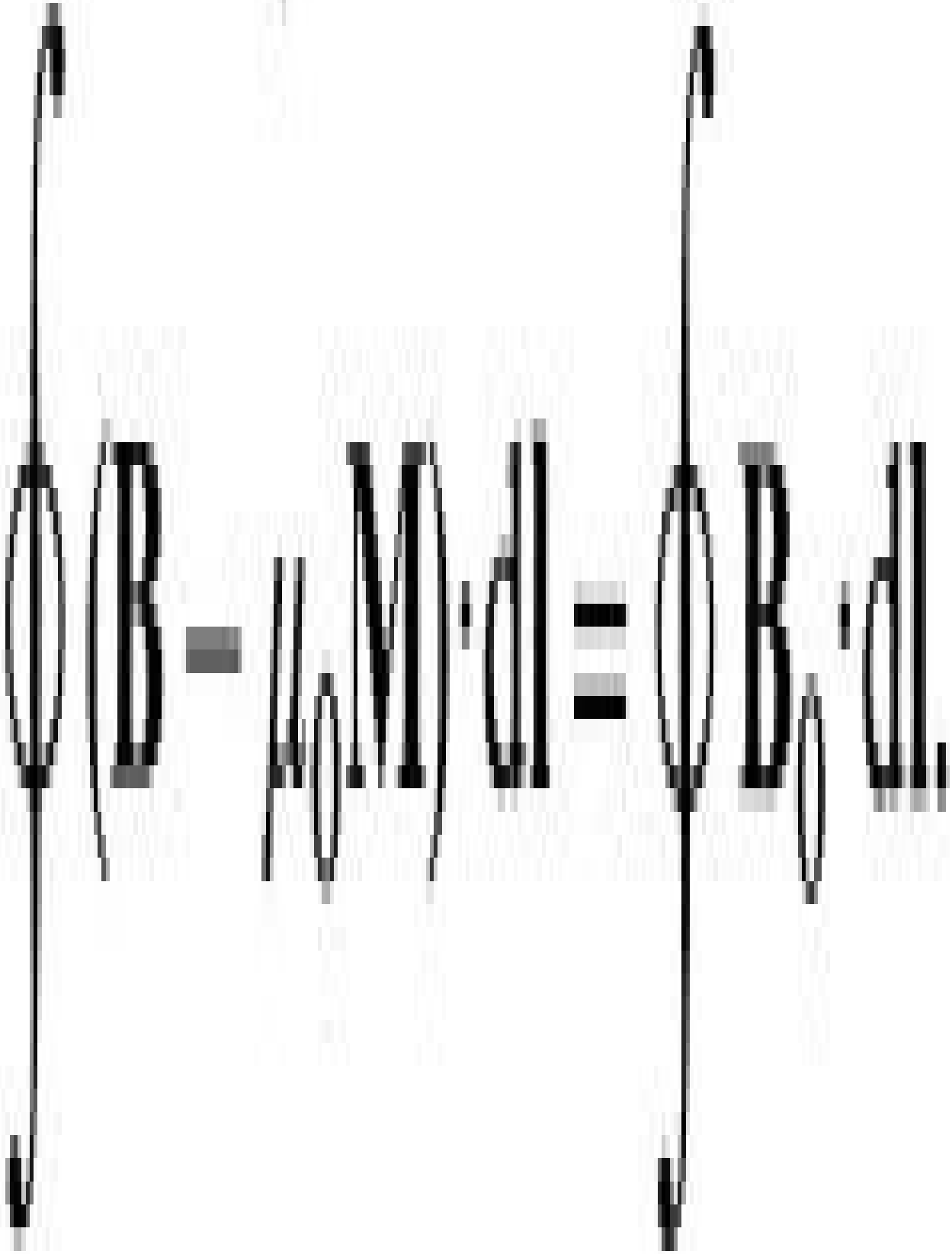

A
B
D
C
I

where B0 and BM have their usual meanings. From Equation (5.15)

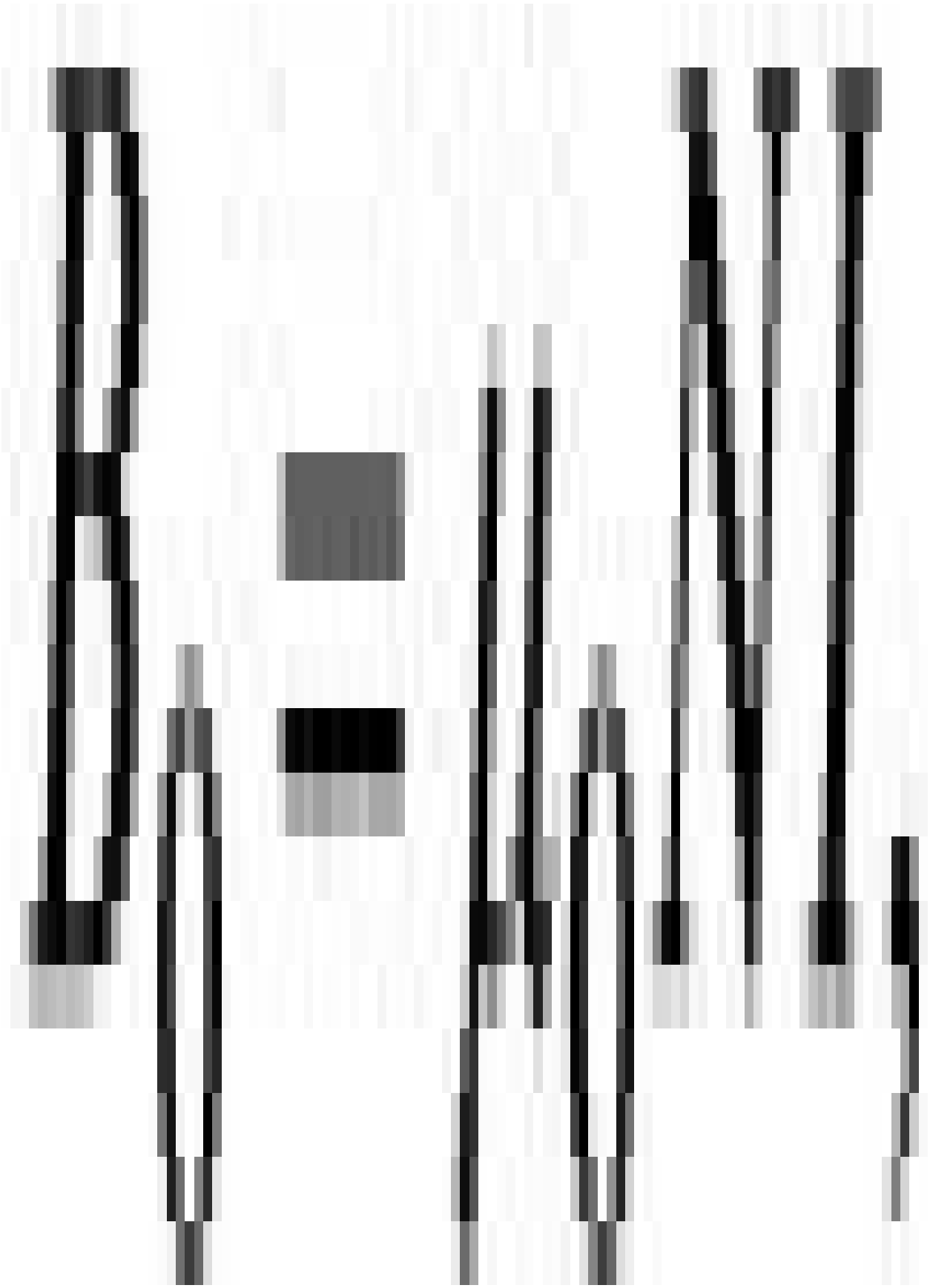
where M is the magnetization, which is zero outside the rod. The line integral is thus equal to



and rearranging



From equation (5.14)

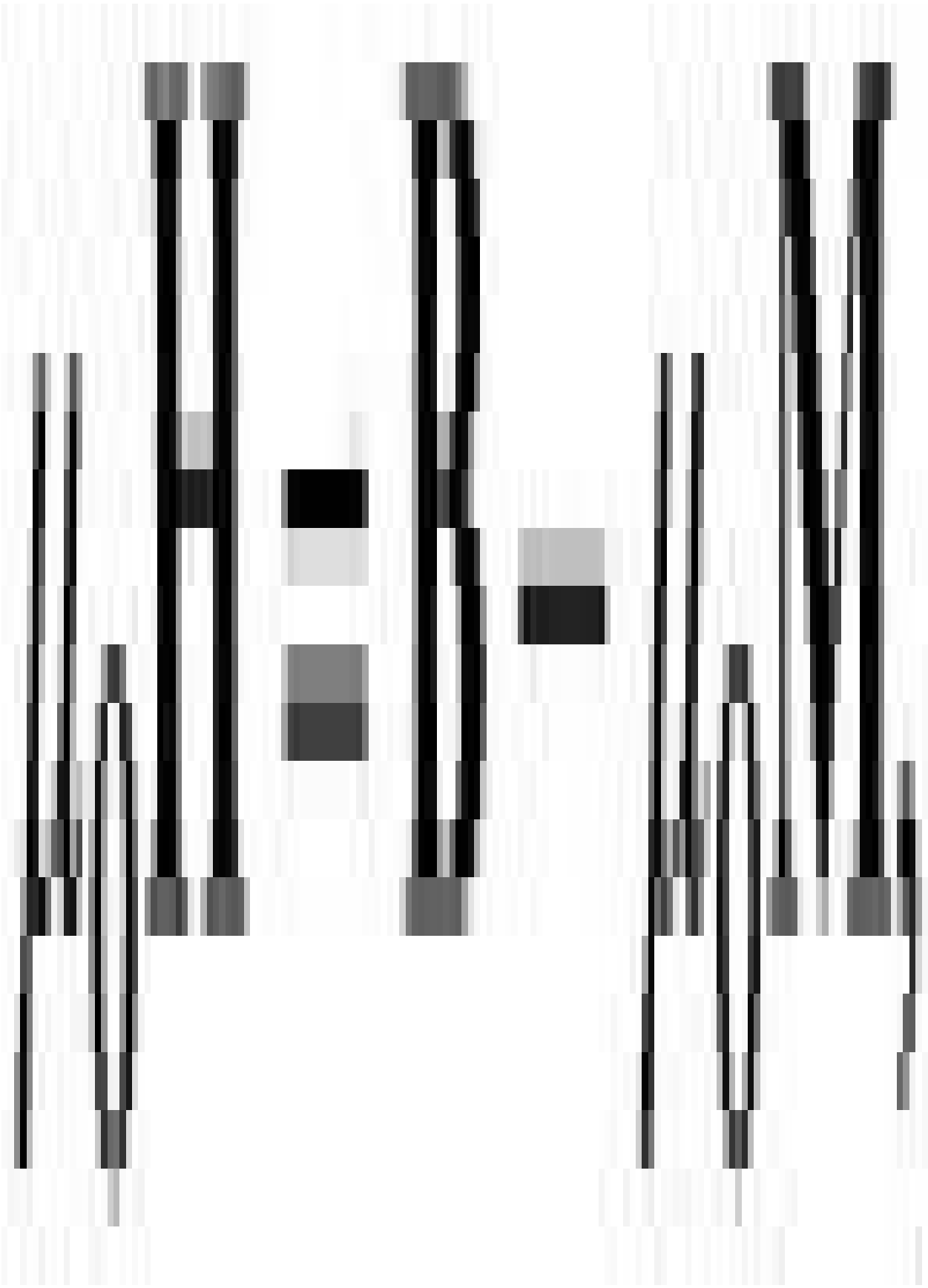


where N is the number of turns per unit length of the solenoid and I the current through the solenoid coils.

We now distinguish between free currents If due to the motion of free charges, such as electrons in wires, and magnetization currents. The total free current flowing through the area ABCD is If = NIl, and hence

(5.26)

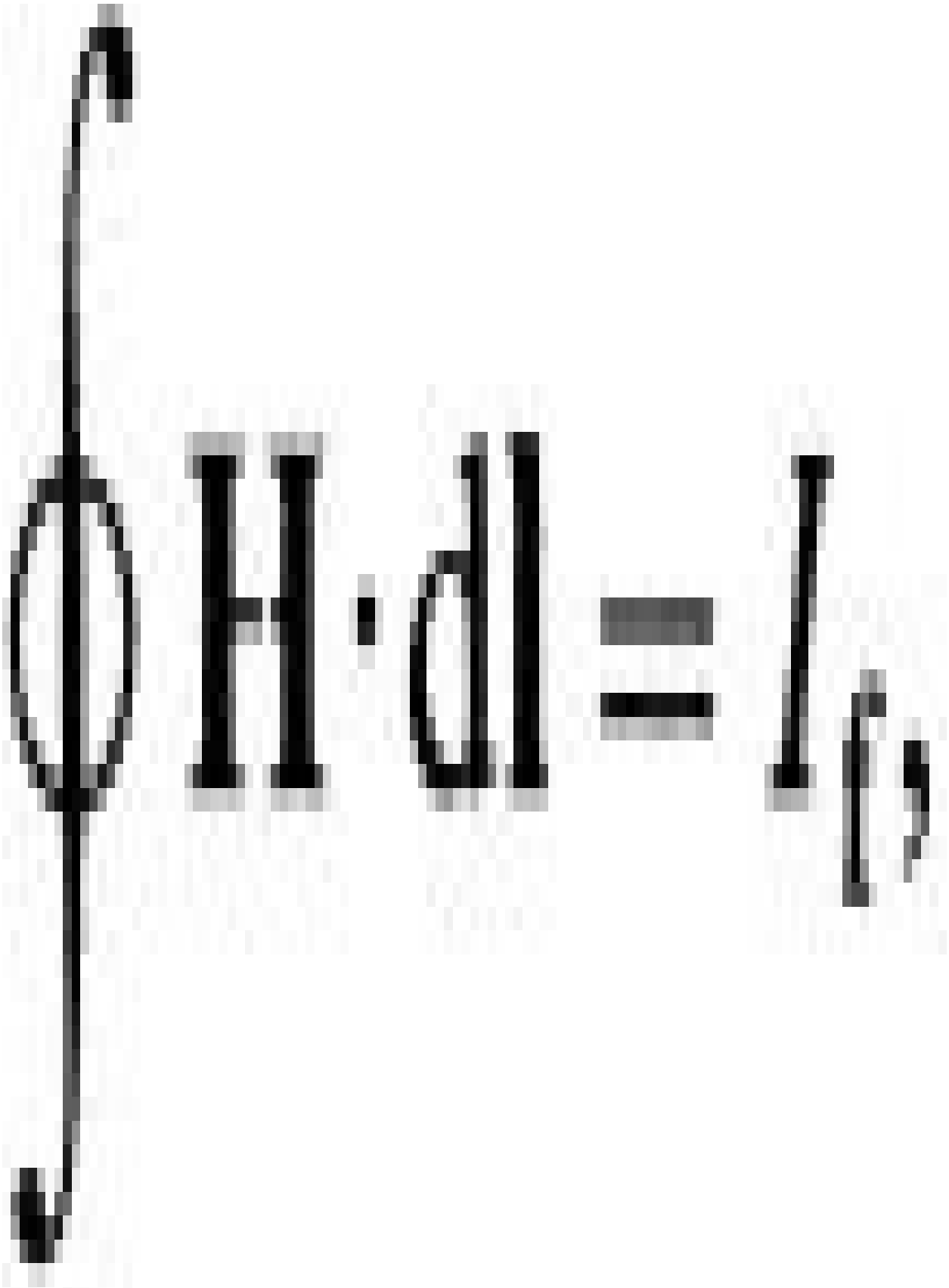
We now define the new auxiliary field vector, the magnetic intensity H, to be given at each point by the equation



or

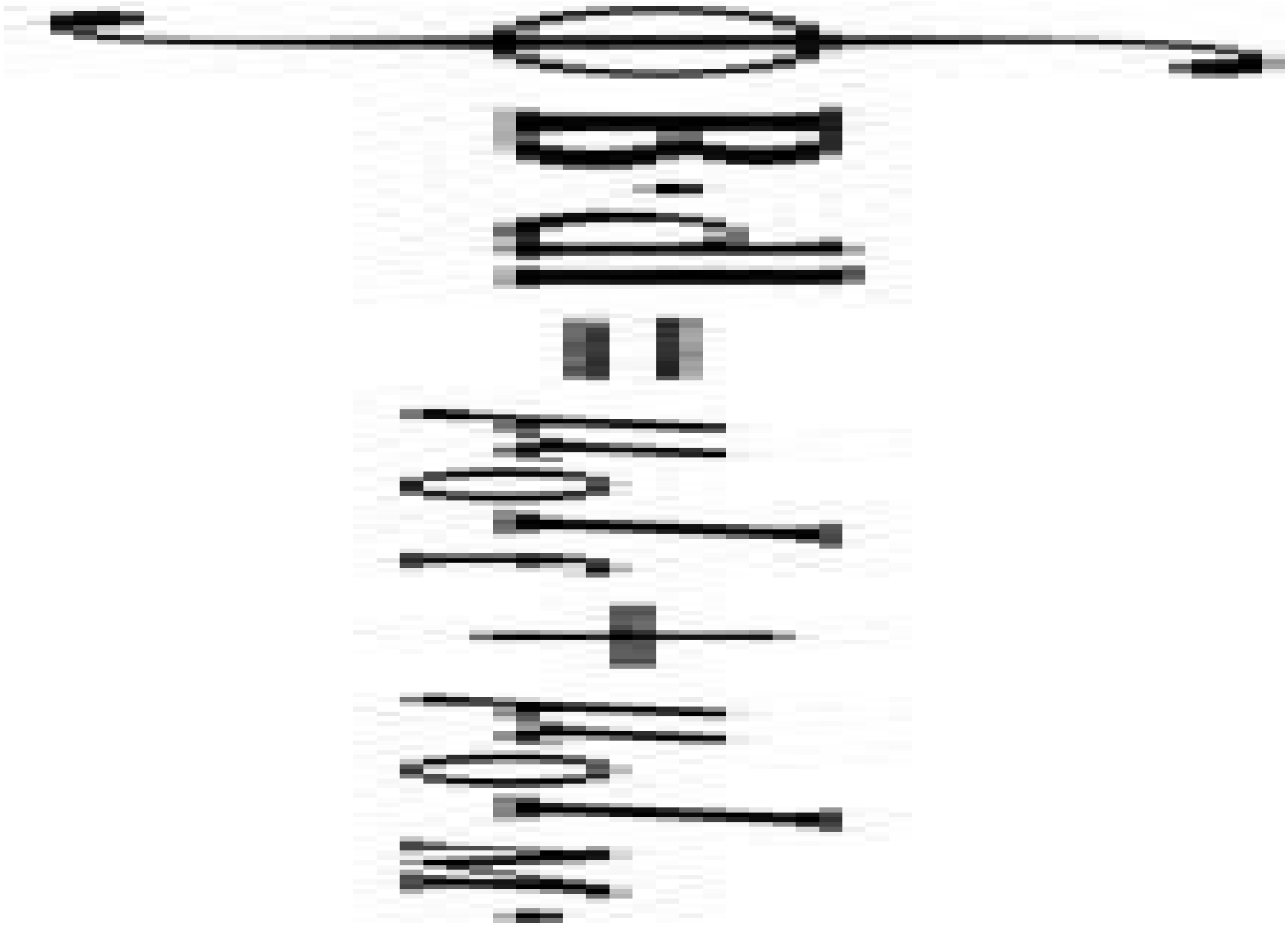
(5.27)

where B and M are the magnetic induction field and the magnetization at the point. In terms of the magnetic intensity, Equation (5.26) becomes



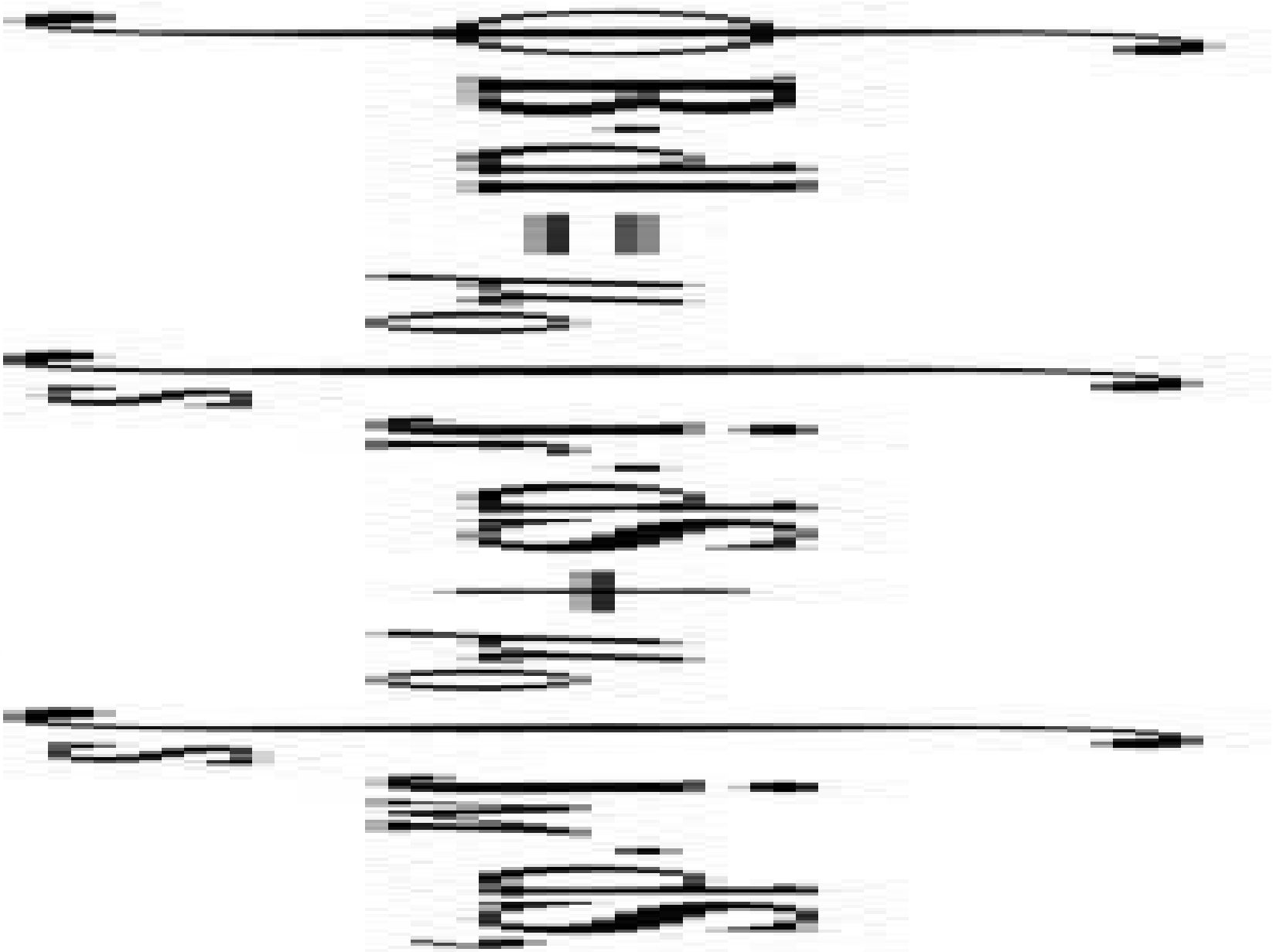
and we see that the line integral of the field H over the closed path depends only on the total current due to the motion of free charges which pass through the area defined by the path. The dimensions of magnetic intensity are current divided by length, the same as the dimensions of magnetization. In SI units magnetic intensity is measured in amperes per metre.

The magnetic field H may be introduced in a more general way, beginning with Ampère's law (4.31). The total current through a closed loop may include contributions from magnetization currents IM as well as from currents If due to the motion of free charges, and we have

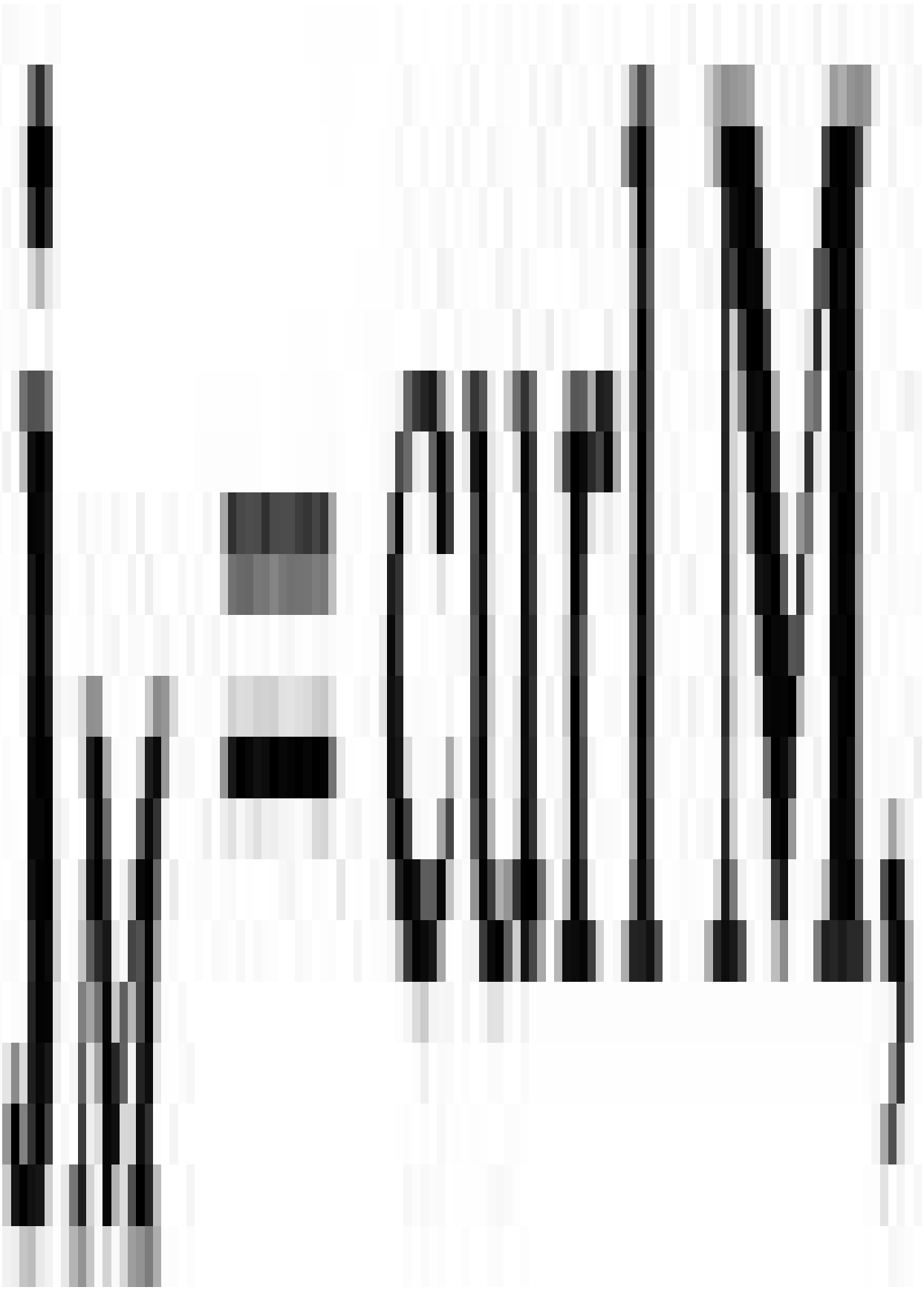


Writing this equation in terms of the corresponding current densities jf and jM

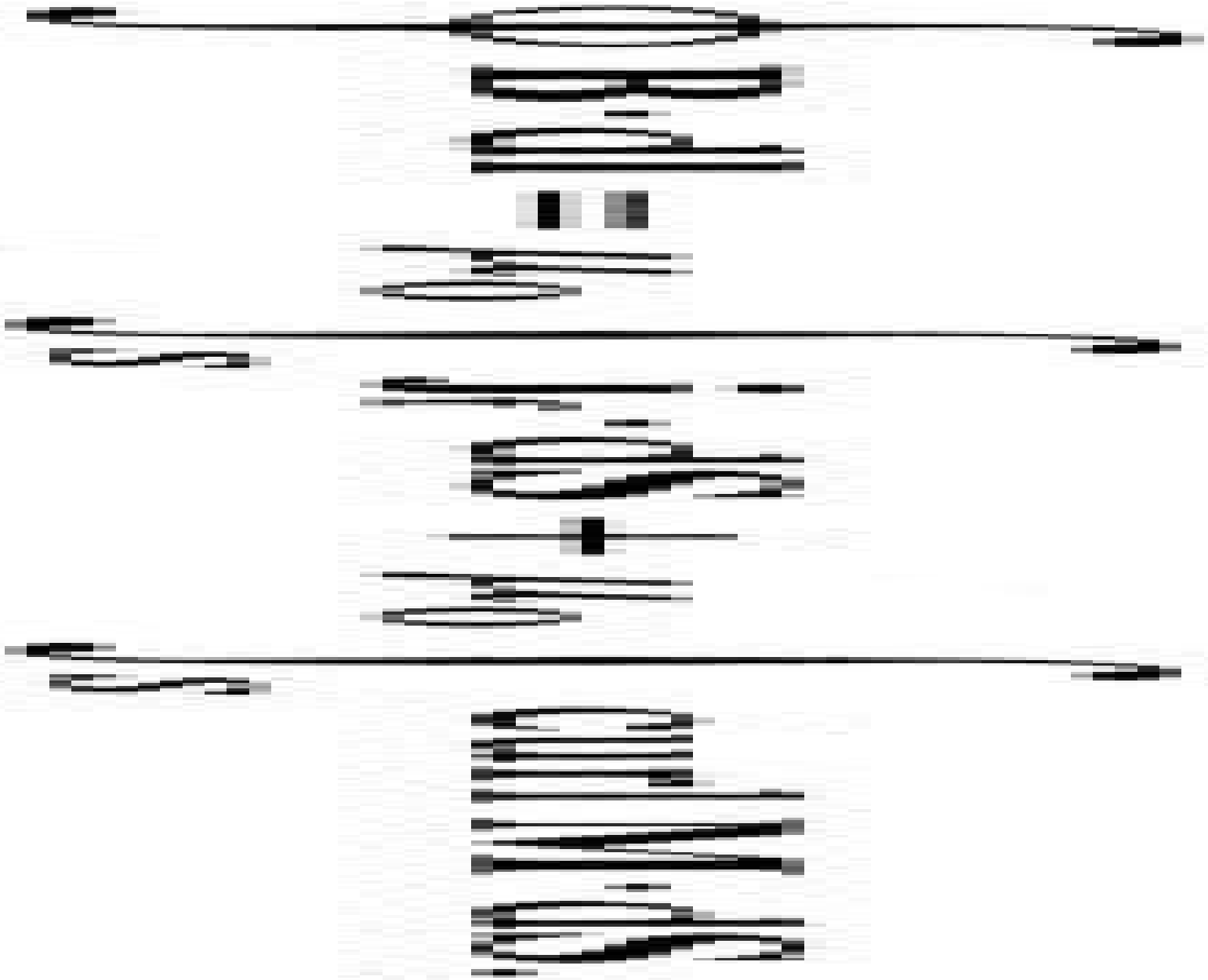
(5.28)



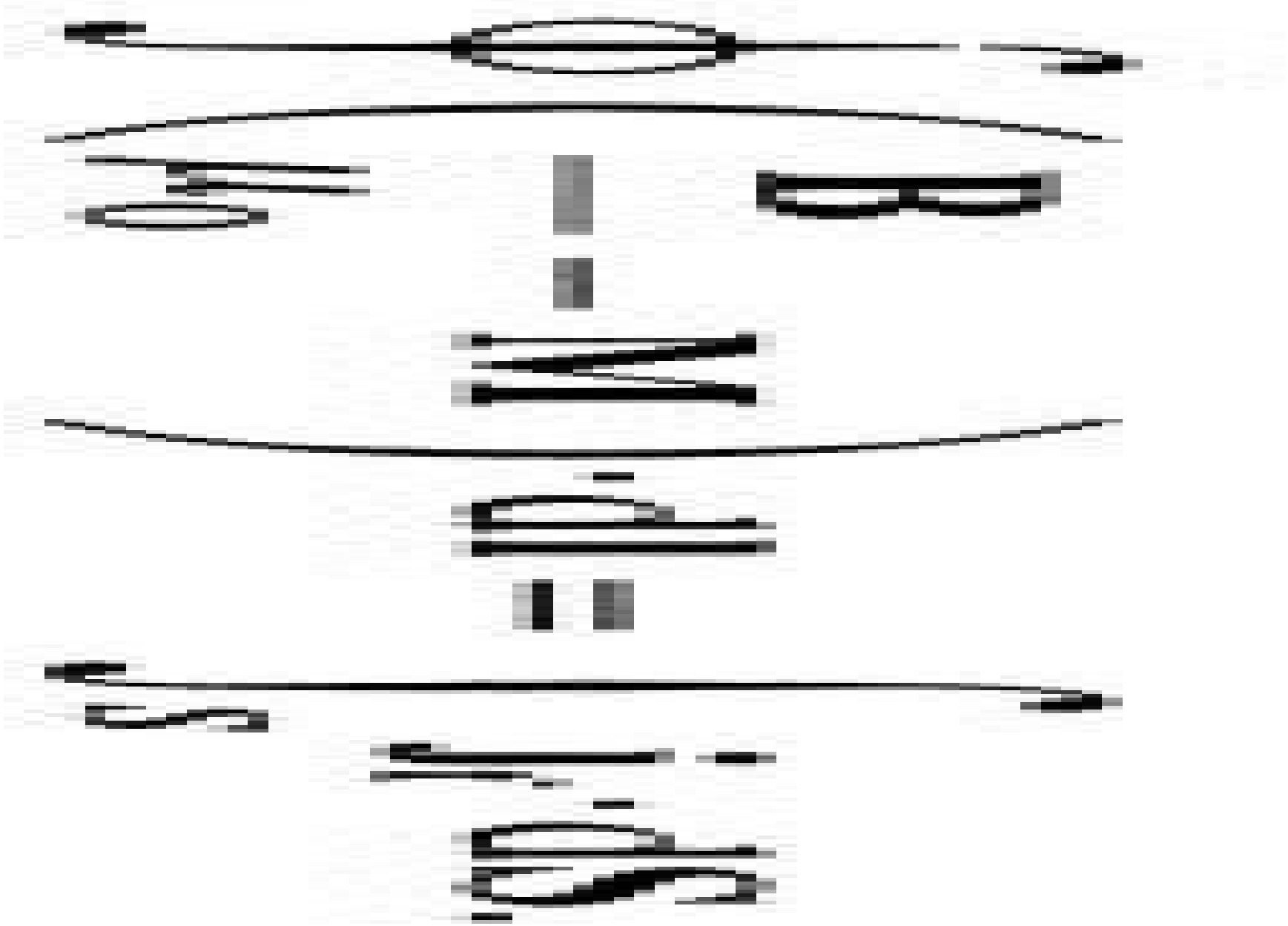
where S is any surface whose perimeter is the path of the line integral. From Equation (5.19)



where M is the magnetization, and Equation (5.28) can be written



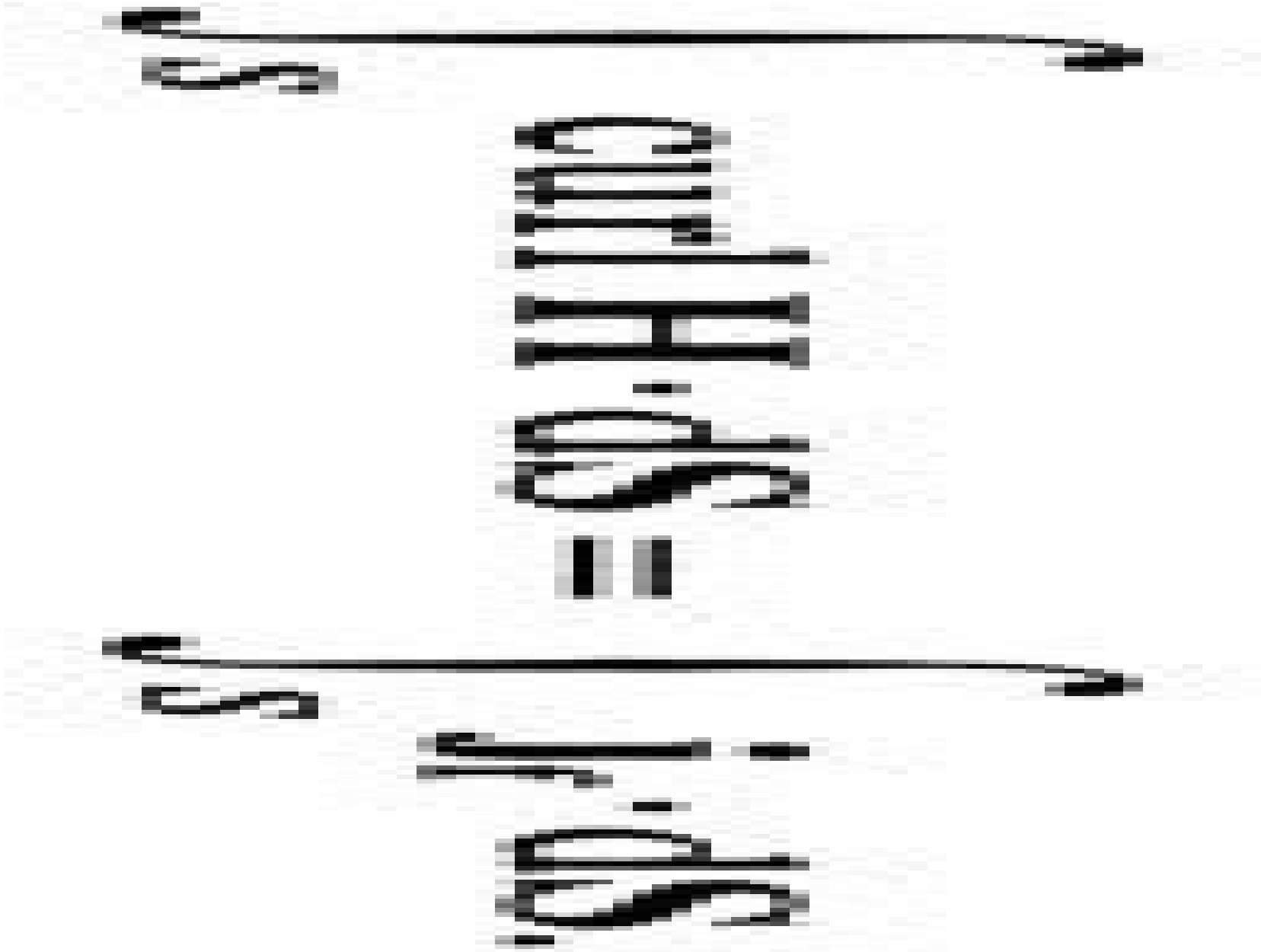
The second term on the right-hand side of this equation can be transformed into a line integral using Stokes' theorem (4.45), and after rearranging we have



In terms of the magnetic field H this equation becomes

(5.29)

This expression relating the line integral of the field H over a closed path to the ‘free’ current through the loop is the integral expression of Ampère’s law for the field H. The differential form of the law may be obtained by rewriting Equation (5.29), using Stokes’ theorem, in the form



This equation is true for any surface S, hence at every point

(5.30)

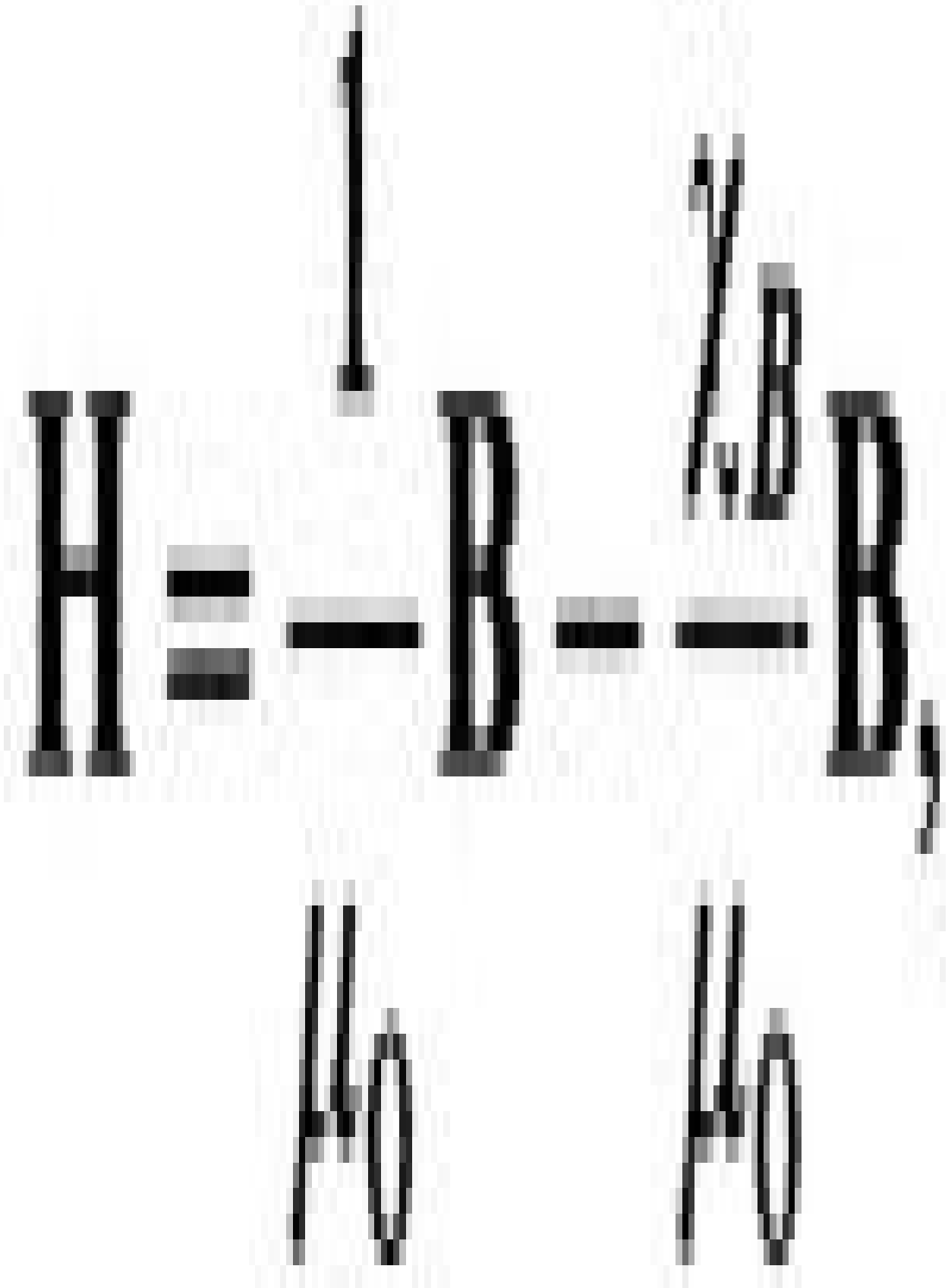
The curl of the magnetic intensity is everywhere equal to the current density due to the motion of free charges.

In the following sections of this book we shall frequently omit the subscript f on currents and current densities when it is clear from the context that the currents referred to arise from the motion of free charges. We will usually use the subscript ‘f’ only when we wish to emphasize the difference between ‘free’ currents and magnetization currents.

For diamagnetic or paramagnetic materials which are isotropic and homogeneous the magnetization is proportional to the magnetic field B, as expressed in Equation (5.23),

(5.31)

This relationship is also approximately valid for ferromagnetic materials over small ranges of the field B. For these materials the susceptibilities are much higher than for non-ferromagnetic materials, and vary with magnetic field. Equation (5.31) enables us to write the magnetic intensity in terms of the field B and the susceptibility. Substitution of the expression (5.31) for the magnetization into Equation (5.27) gives



from which we obtain the equation

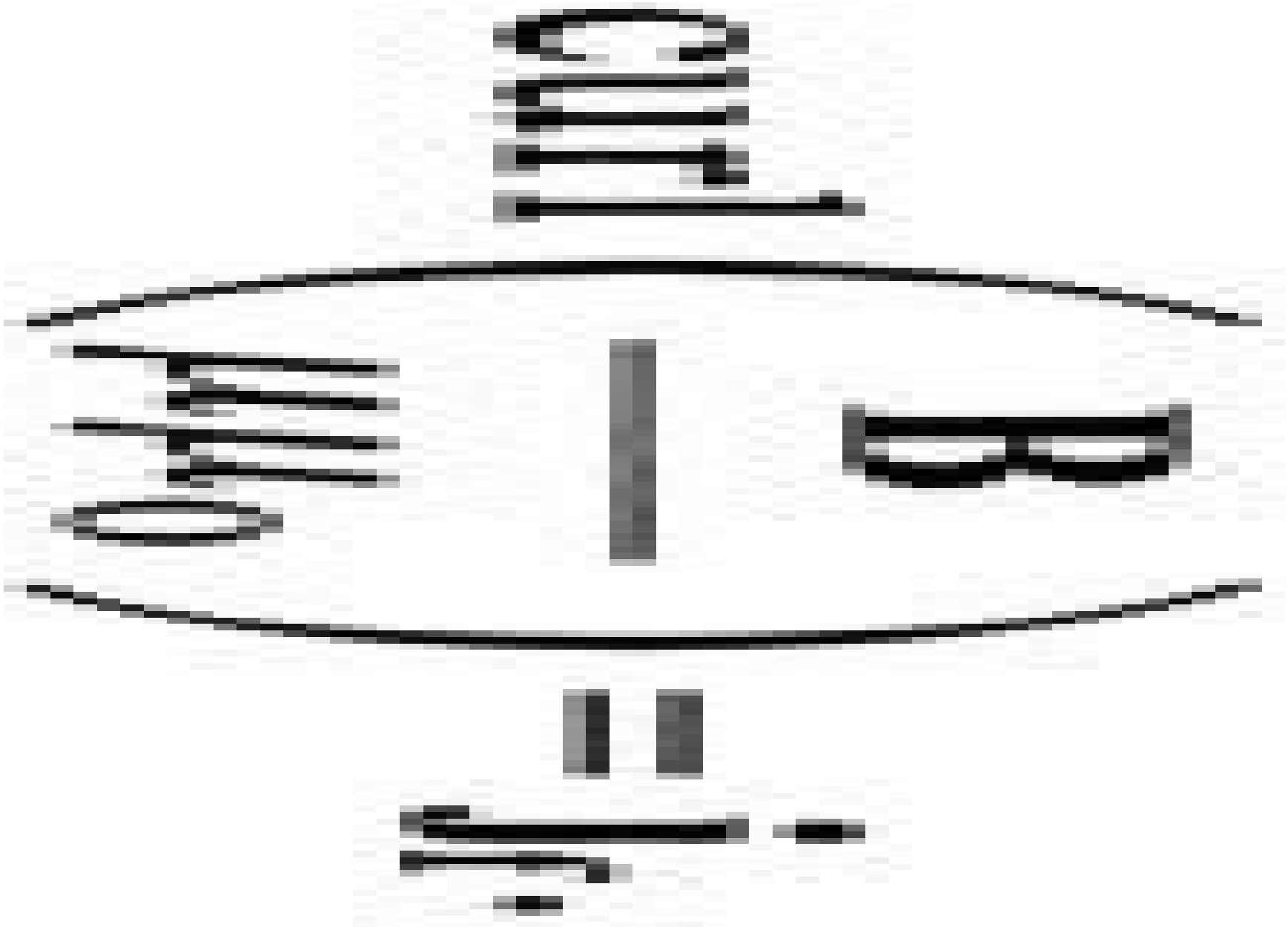
(5.32)

where

(5.33)

The constant μ is dimensionless and is called the relative permeability of the material. For non-ferromagnetic materials the relative permeability is very nearly unity, but for ferromagnetic materials it may have large values, typically around 1000, and like the susceptibility varies with the magnetic field.

If we insert expression 5.32 for the field H into Equation (5,30) we obtain

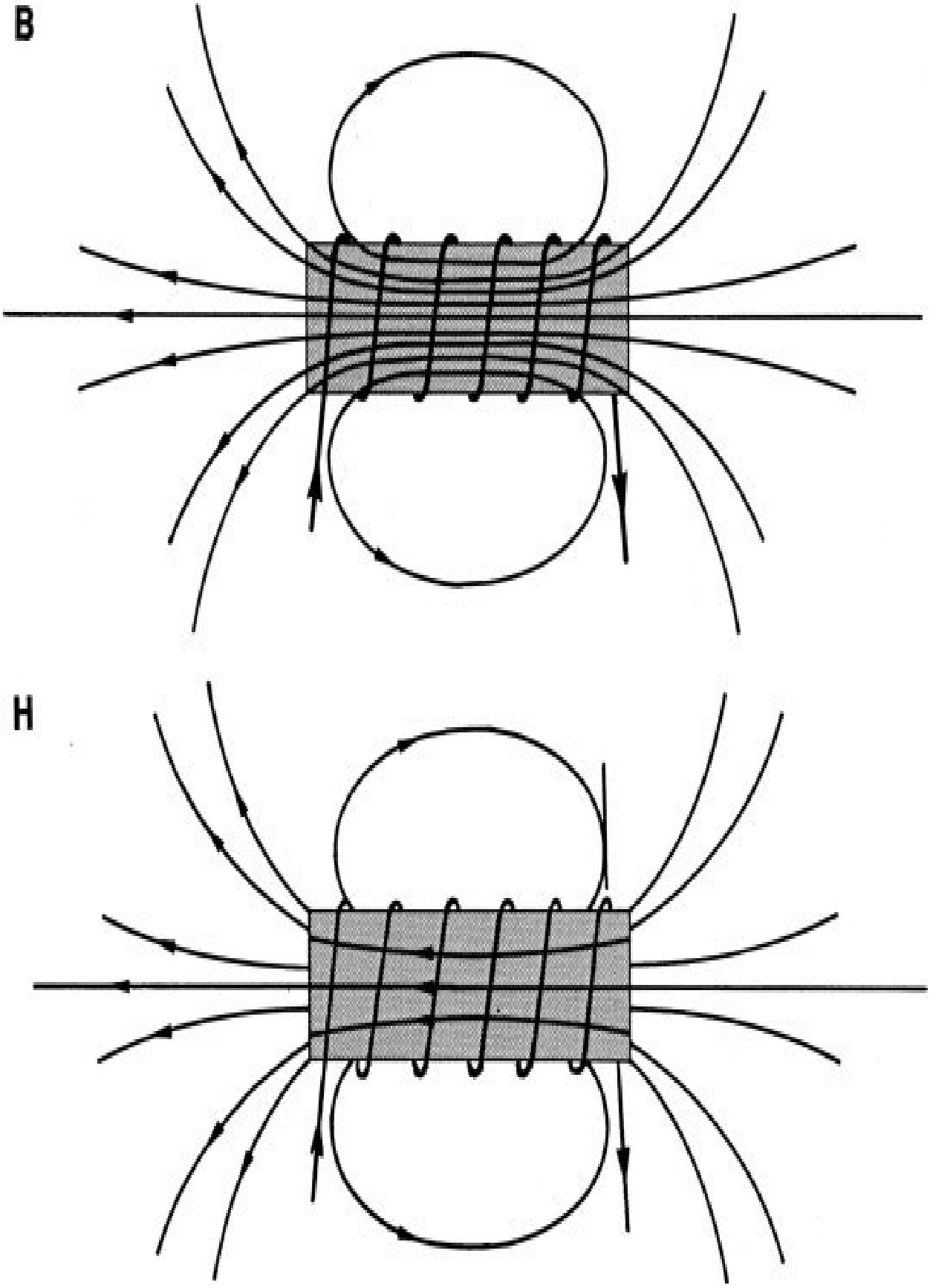


If μ is constant everywhere, this reduces to the equation

(5.34)

which is similar in form to Equation (4.47) except for the factor μ. Thus for a given distribution of currents flowing in an infinite medium of relative permeability μ, the magnetic field B everywhere is increased by the factor μ over the field produced by the same distribution of currents in free space. The situation can be compared with the analogous one in electrostatics. The electric field due to a system of charges at rest in an infinite medium of relative permittivity ε is reduced everywhere by the factor ε compared with the field due to the same system of charges in free space.

Figure 5.10. The lines of the field B (upper diagram) and H (lower diagram) for a short solenoid with a paramagnetic rod inside. The relative permeability of the hypothetical paramagnetic material has been taken to be about three in order to clarify the difference between the fields.


B
H

Unlike the magnetic field B, the field H has sources, and the lines of the field H are not continuous. Taking the divergence of each side of Equation (5.27),

(5.35)

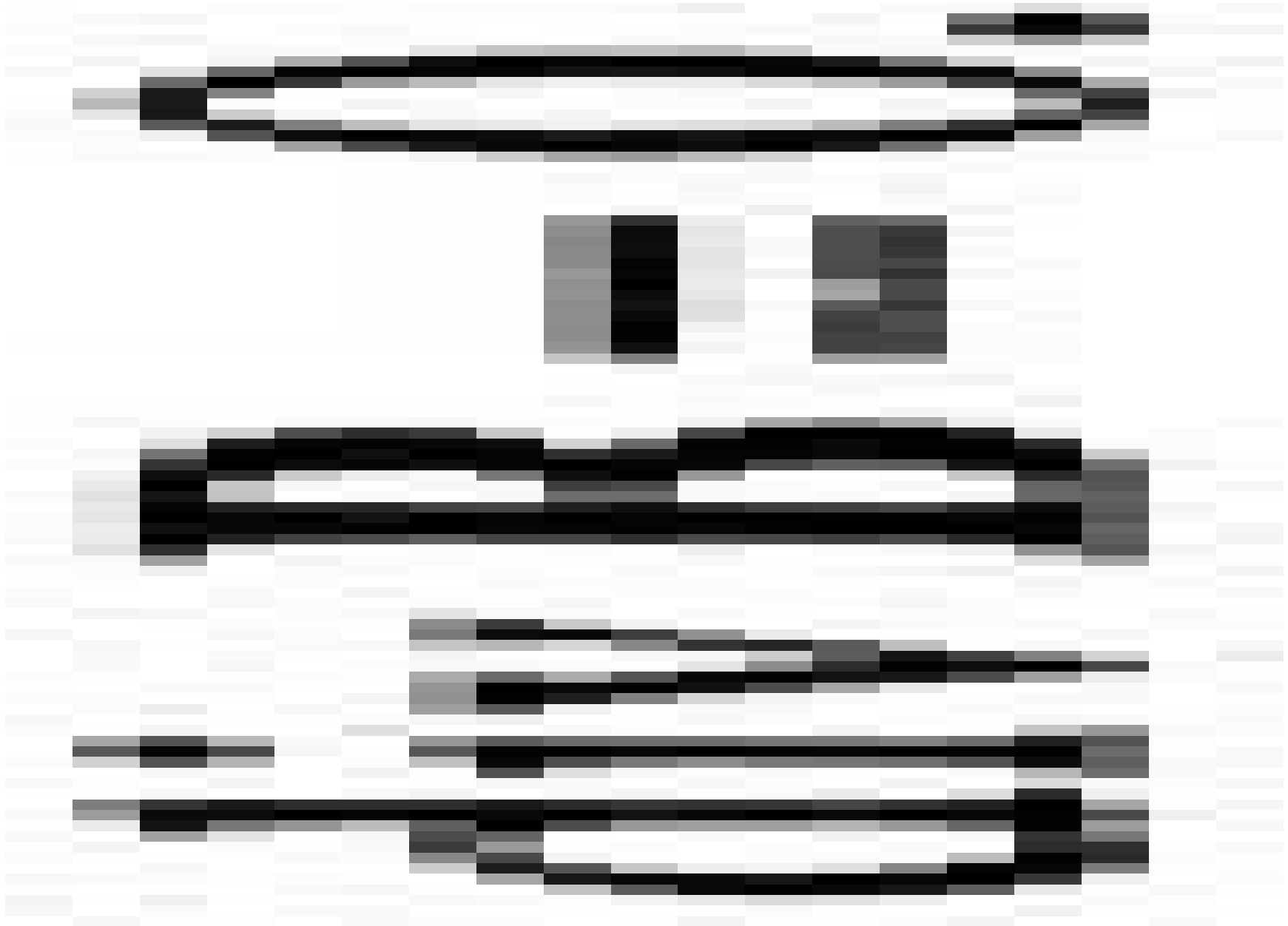
The lines of the combined field (H + M) are thus continuous and at boundaries between different media, where the magnetization M changes abruptly, the field H must change to compensate. For example Figure 5.10 shows the lines of the fields B and H in and around a magnetized paramagnetic rod. At the ends of the rod, where the magnetization changes discontinuously, more lines of the field H occur outside than inside. Over a small volume enclosing part of the left end of the rod more lines of the magnetization flow in than out, and more lines of magnetic intensity flow out than in.

If the free charge current distribution is specified, and the geometry of magnetic media present is known, Equation (5.30) can in principle be solved to give the magnetic intensity everywhere. Such a situation occurs in the example with which we began this section in which we wanted to know the field near the end of a rod inserted in a solenoid. At the boundaries between different media there will be difficulties, due to the fact that the field H has sources there. However, we will find that at these boundaries the fields H and B, related by the equation B = μμ0H must obey certain conditions. These enable one to match the solutions to Equation (5.30) in different regions of space containing materials of constant permeabilities. If we can find fields H that satisfy Equation (5.30) in each of the regions of space, and which obey the boundary conditions, these fields make up the solution we are looking for, since the uniqueness theorem is valid for magnetostatics as well as for electrostatics for materials which have a constant permeability.

5.3.2 The boundary conditions on the fields B and H

In this section we will derive the conditions that the field vectors B and H must obey at the boundaries between different media.

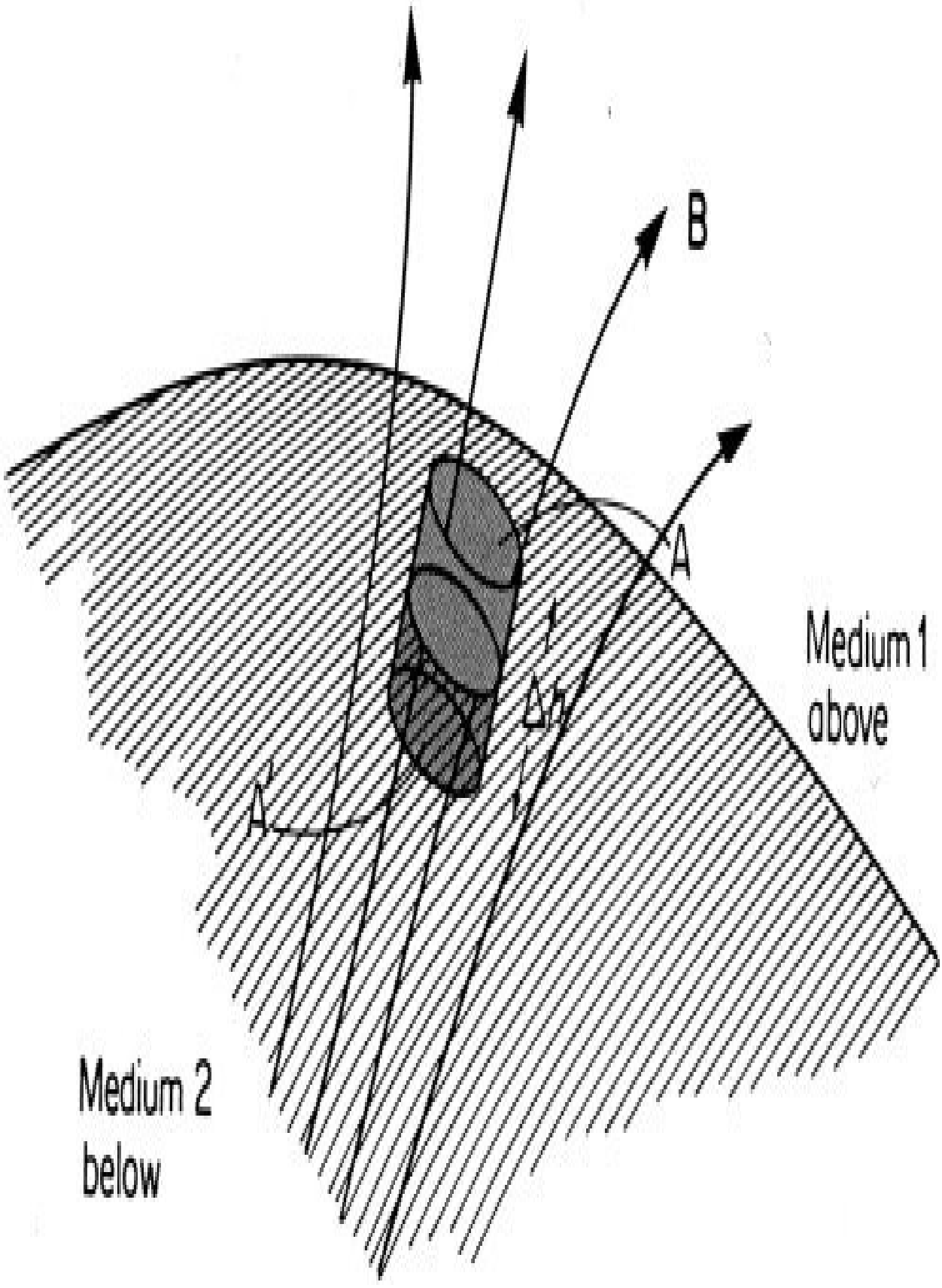
We can derive a condition on the field B based on the equation



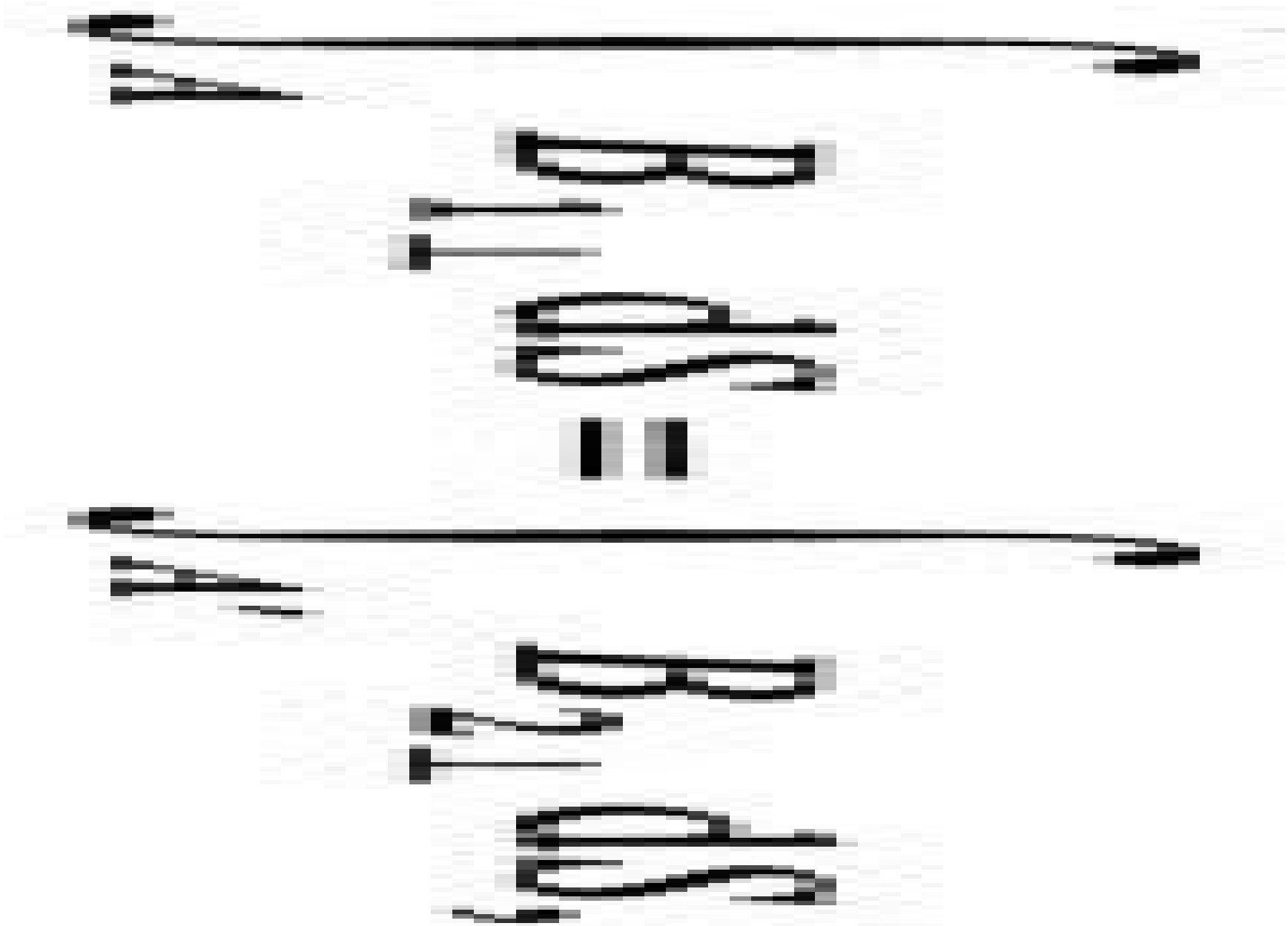
This equation means that the flux of the field B out of any closed surface is zero. The field lines are continuous and therefore as many lines enter the volume enclosed by the surface as leave it. Consider the closed surface shown in Figure 5.11. This is the surface of a small disc of height Δh that sits astride the boundary between two magnetic media. As we make the height Δh tend to zero, the total flux of magnetic induction over the surface of the disc has contributions from the top and bottom ends only, and

(5.36)

Figure 5.11. The boundary condition on the magnetic field B.


B
A
Medium 1
above
Δh
A'
Medium 2
below

Now $B \cdot dS = \pm B_{\perp}\, dS$, where $B_{\perp}$ is the component of the field B normal to the plane of the surface dS. Equation (5.36) becomes



where $B1\perp$ is the component of the field B in medium 1 normal to the interface, and $B2\perp$ is the component in medium 2 normal to the interface. Note that the signs are correct because dS is a vector perpendicular to the plane of dS and in a direction out of the volume enclosed by the surface. The above equation is true whatever the size of the (equal) areas A and A'. The only way this can be so is for the condition

(5.37)

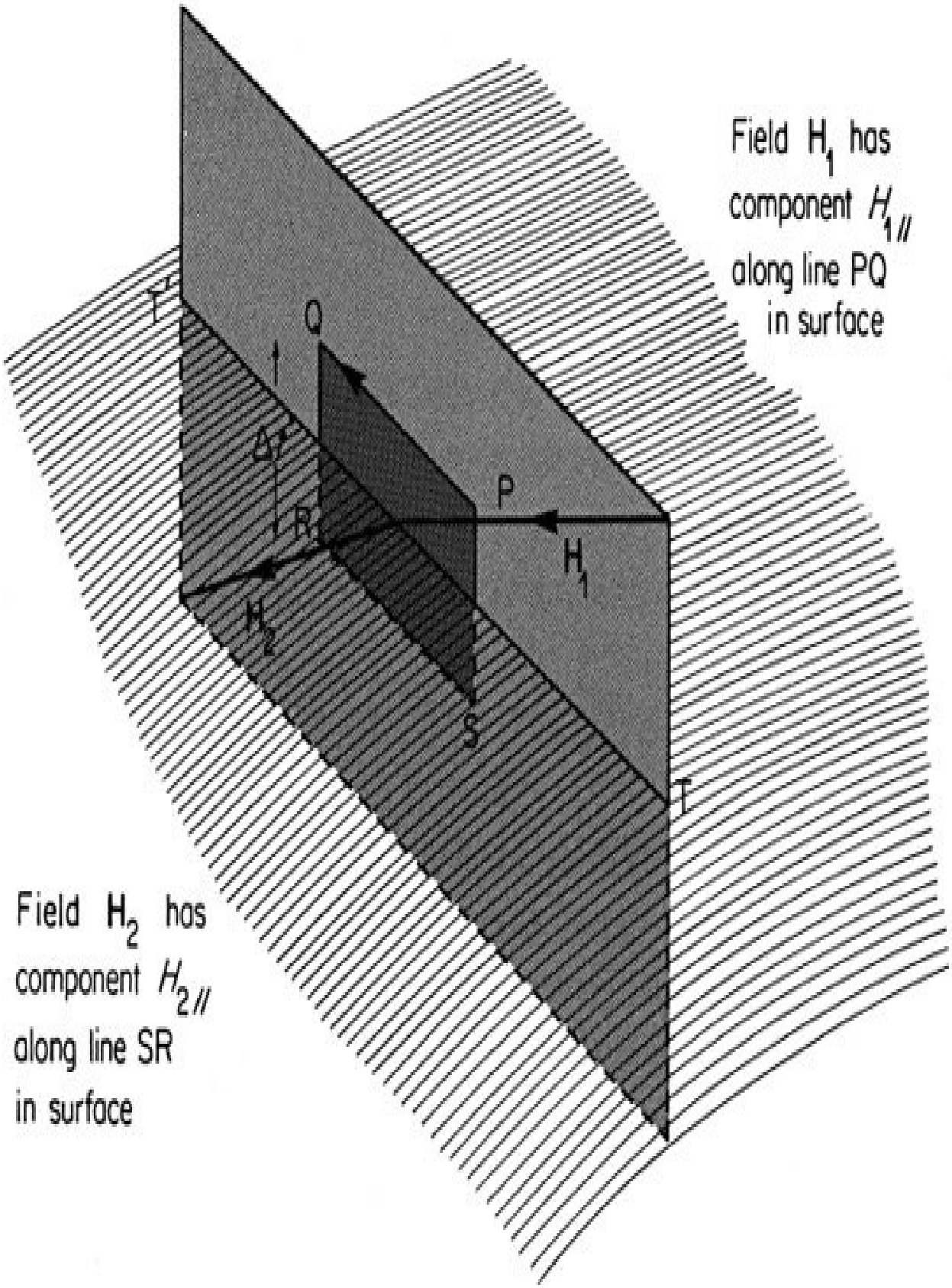
to hold at every point on the surface separating the two media. The normal component of the magnetic field B at every point on the interface between two media is continuous. Note that this is not true for the field H because at the boundary the magnetization M is discontinuous, and the field H has sources.

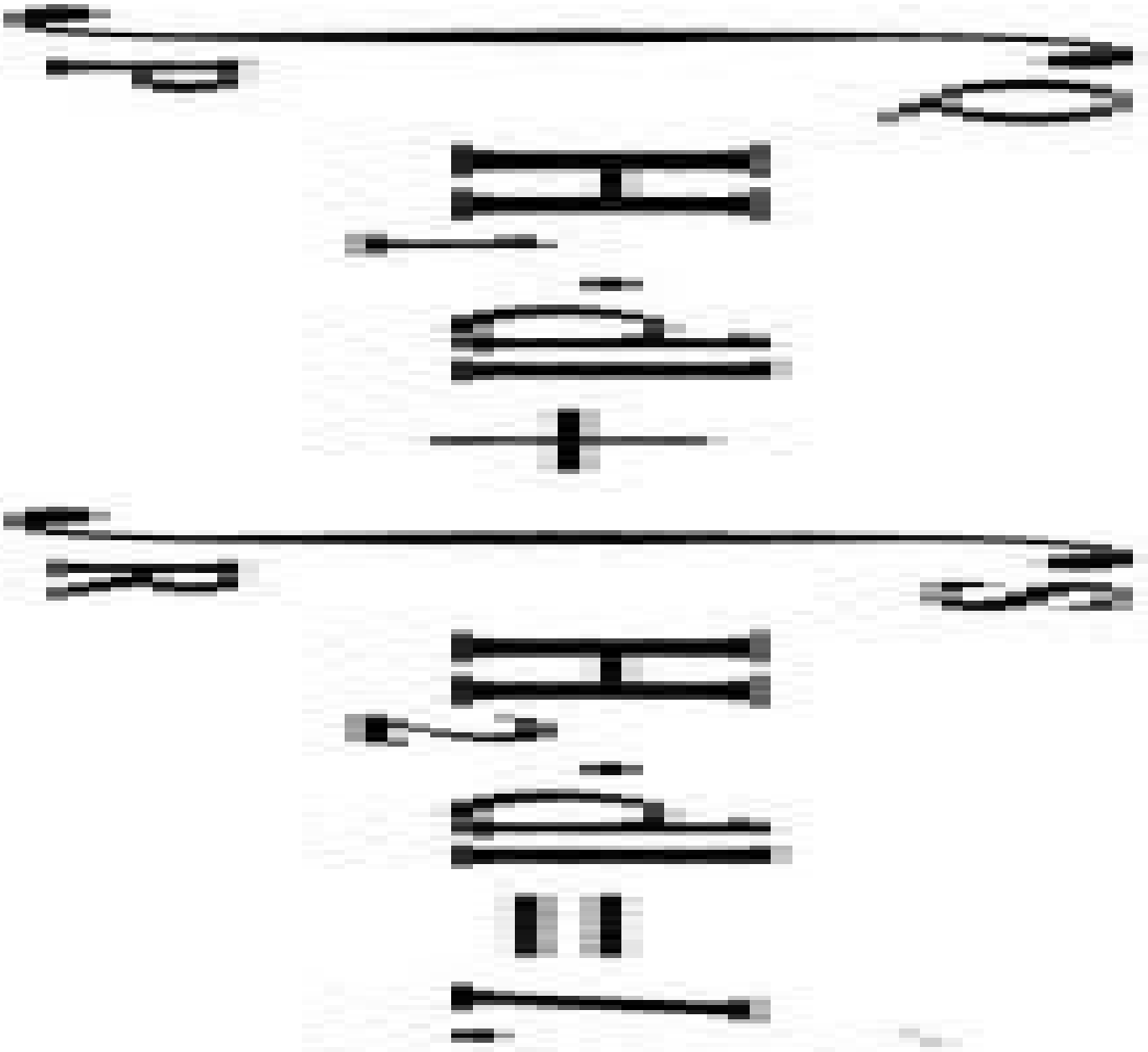
We can derive a condition on the field H based on Equation (5.29).

(5.38)

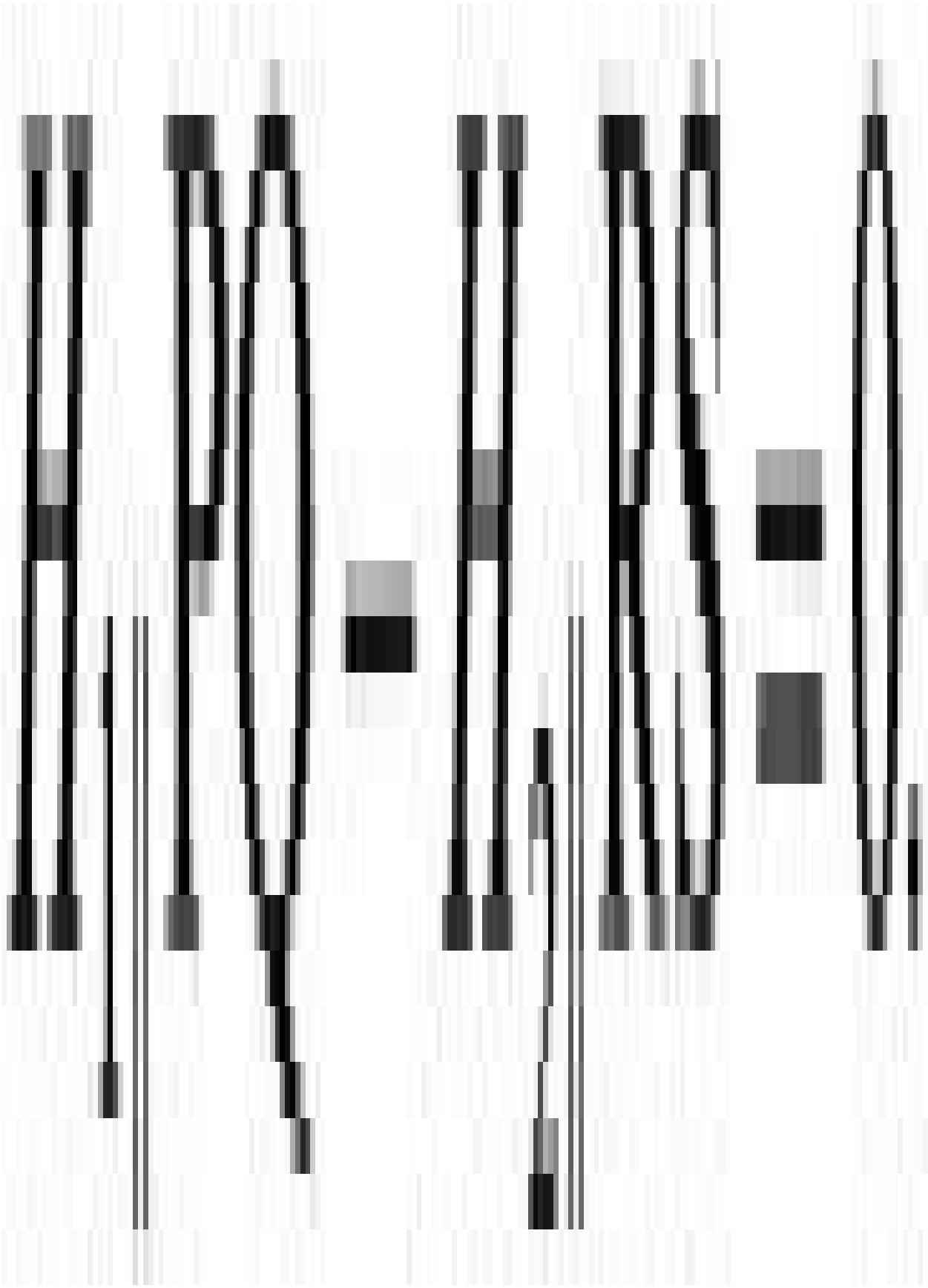
where I is the free charge current that passes in a direction given by the right-hand rule through the area defined by the path over which the line integral is taken. Consider a path PQRS as shown in Figure 5.12, where the lengths PQ and RS follow the contour of the surface on either side of the line TT′ drawn in the interface, and are in the direction of the surface component of the magnetic intensity. The lengths QR and PS lie along normals to the surface and are of length Δt. We will apply Equation (5.38) to this path as the length Δt becomes smaller and smaller, tending to zero, in which case

Figure 5.12. The boundary condition on the magnetic intensity H.


Field H_1 has component $H_{1//}$ along line PQ in surface
T
Q
P
R
S
H_1
H_2
T
Field H_2 has component $H_{2//}$ along line SR in surface



As the area of PQRS tends to zero the current I tends to zero, otherwise infinite amounts of charge would flow through the vanishingly small area PQRS. Infinite current density does not occur in any material, although superconductors can sustain very high values of current density and have essentially zero resistivity up to the limits of their current carrying capability.* The current I in the above equation then reduces to zero, and we have



Here H1 and H2 are the maximum components of the fields H1 and H2 in the surface either side of the interface between the media. (The sides PQ and RS of the rectangle were drawn parallel to the projections of the vectors H1 and H2 on the surface.) The boundary condition on the field H is thus finally that

(5.39)

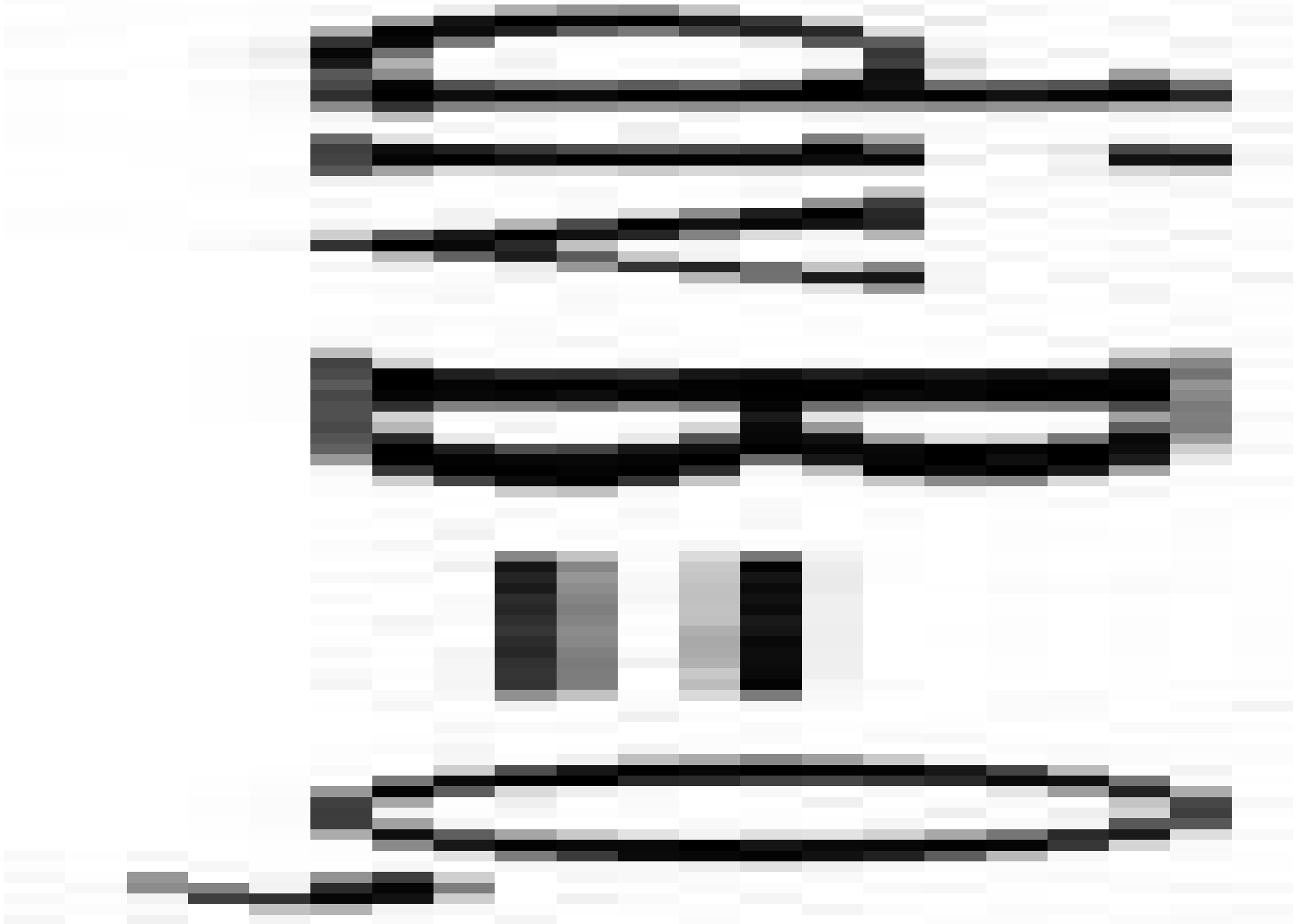
at all points across the boundary between two media.

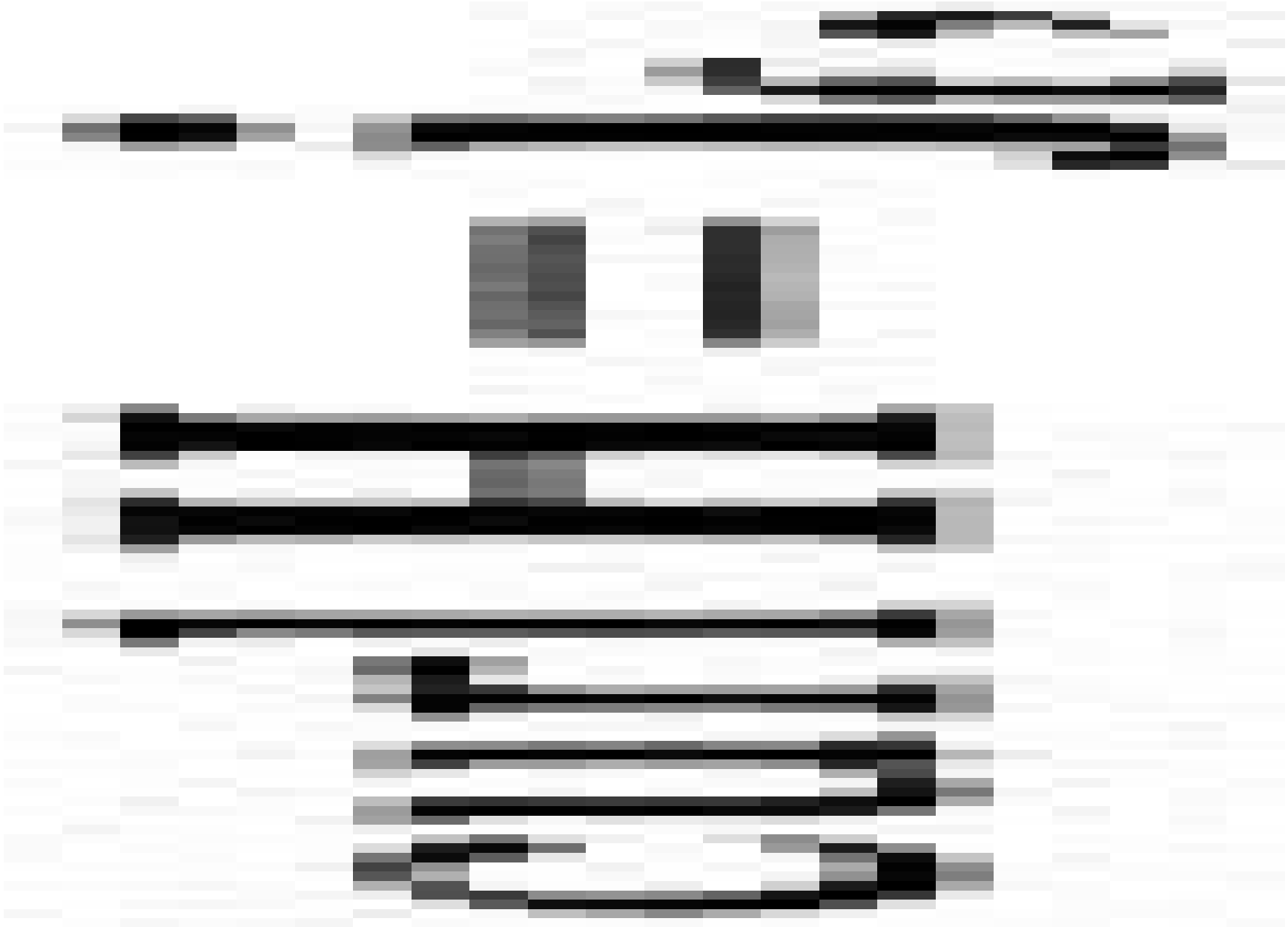
Equations (5.37) and (5.39) are the boundary conditions on the fields B and H we set out to derive at the beginning of this section. They are of great help in the solution of problems in magnetism, as the remainder of this chapter illustrates.

5.4 MAGNETS

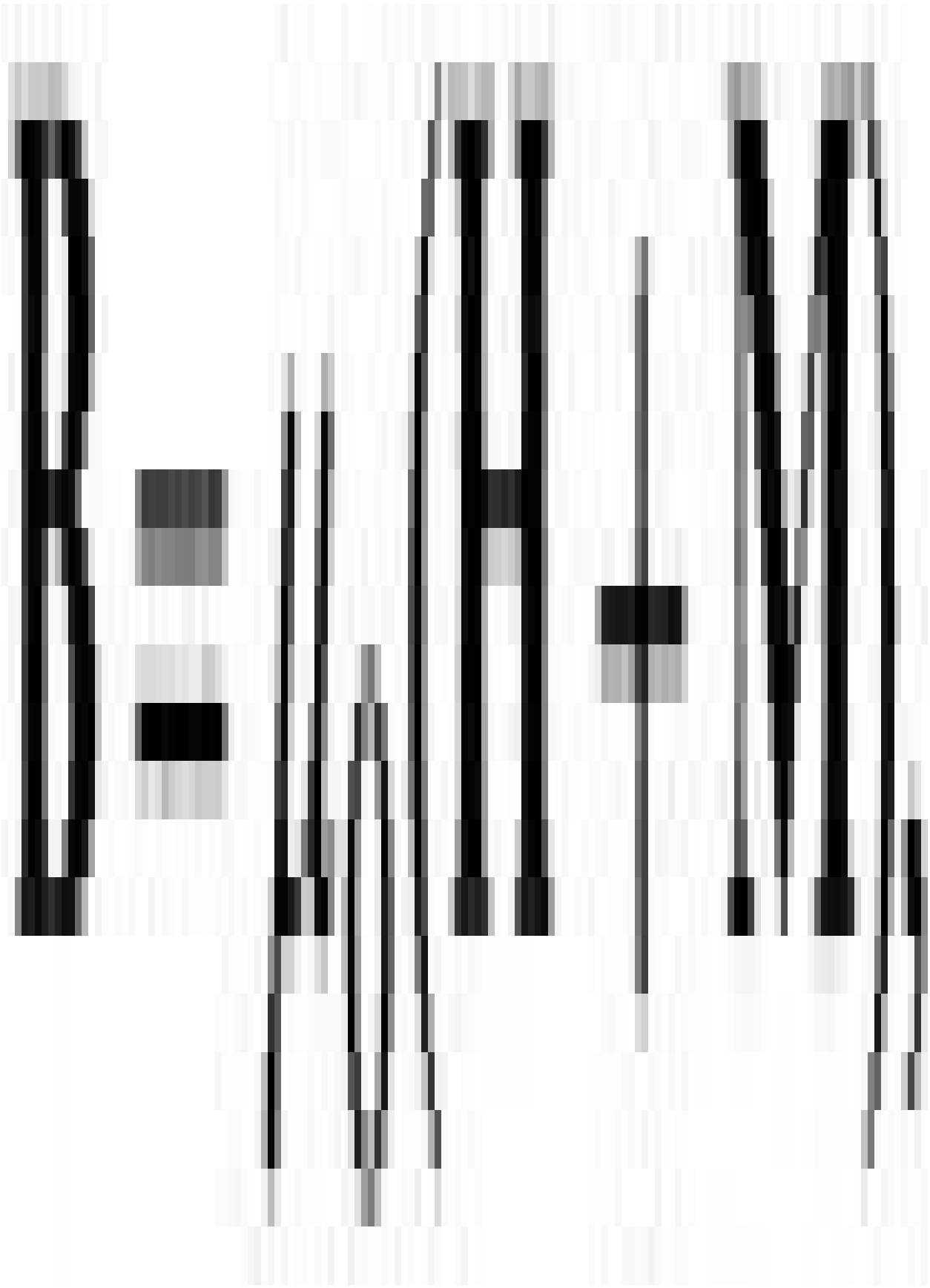
In this section we discuss some aspects of the important practical topic of magnets. Magnets and the use of magnetic materials are ubiquitous in modern technology, and are of great economic importance.

We have shown how the magnetic fields B and H obey certain relations. At all points in a steady magnetic field

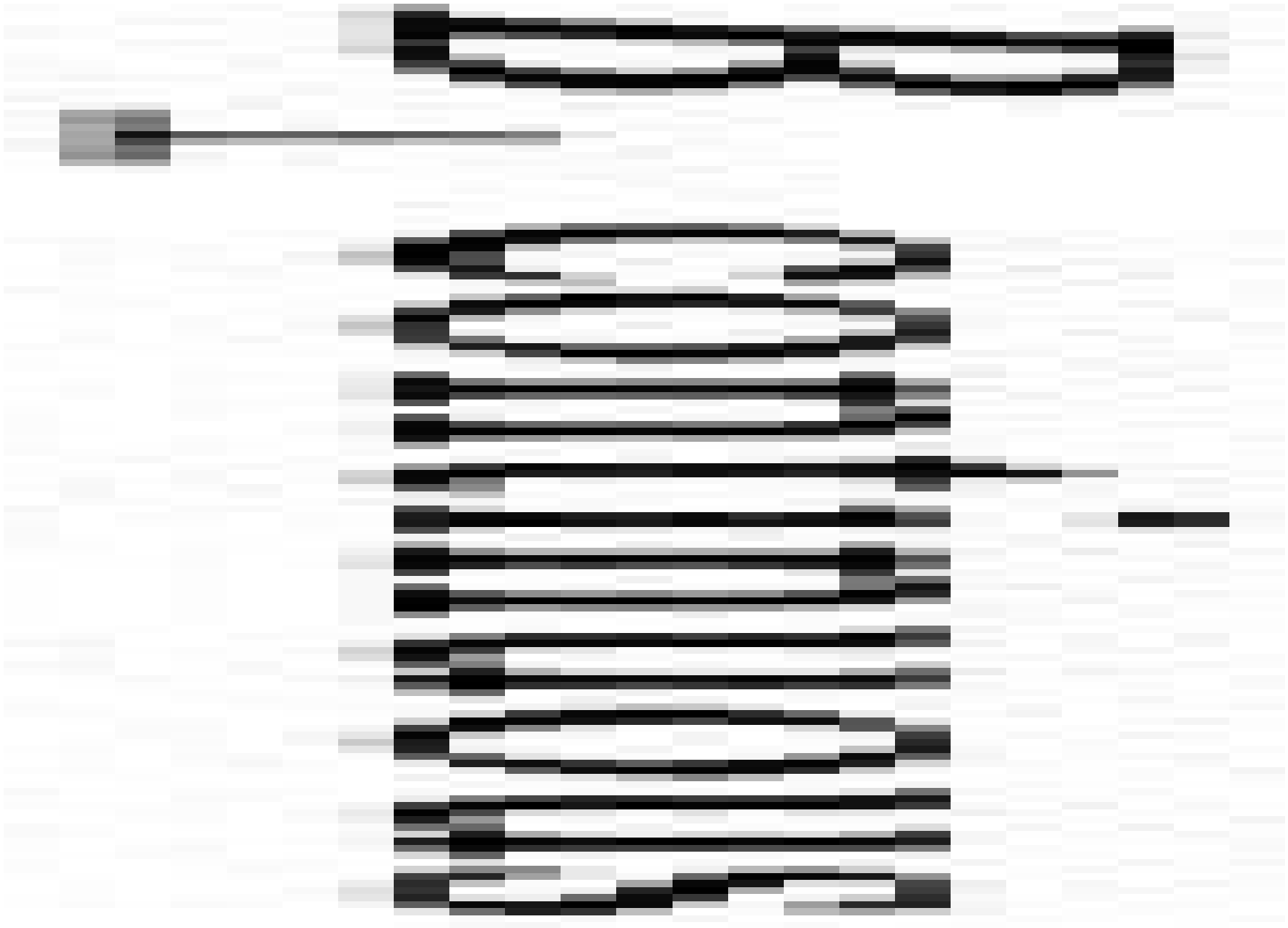




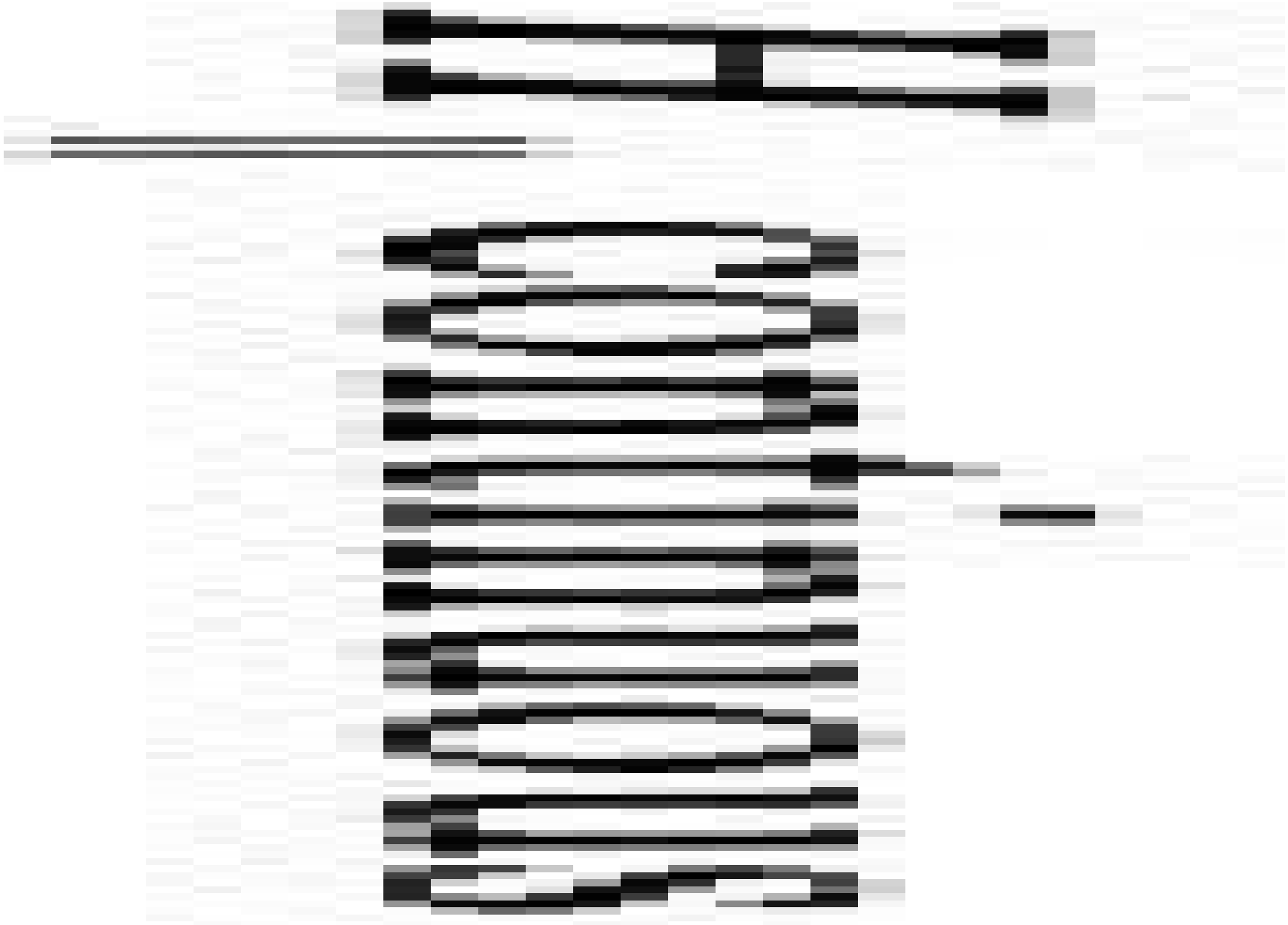
and



where the symbols have their usual meanings. The surface integral ∫sB·dS over any closed surface is zero, and the line integral H·dl around any closed path is equal to the current due to free charges that passes through the area defined by the path in a direction given by the right-hand rule. At the boundary between different media there are discontinuities in the fields, but the latter have to obey certain conditions at these boundaries, viz



and

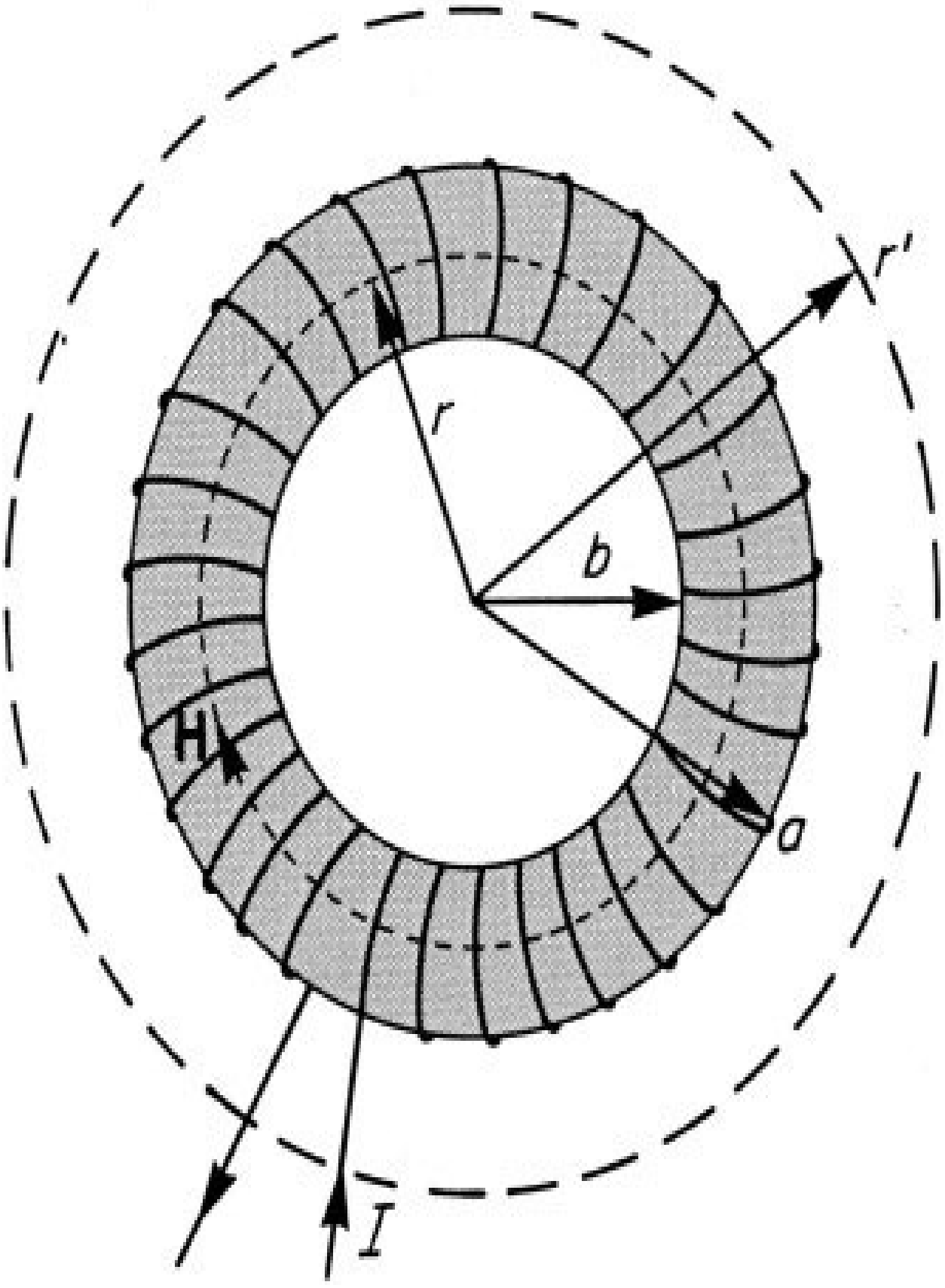


where $B_{\perp}$ is the component of the magnetic field B normal to the interface and H is the component of magnetic intensity H tangential to the interface. We can now solve some simple problems in magnetism.

5.4.1 Electromagnets

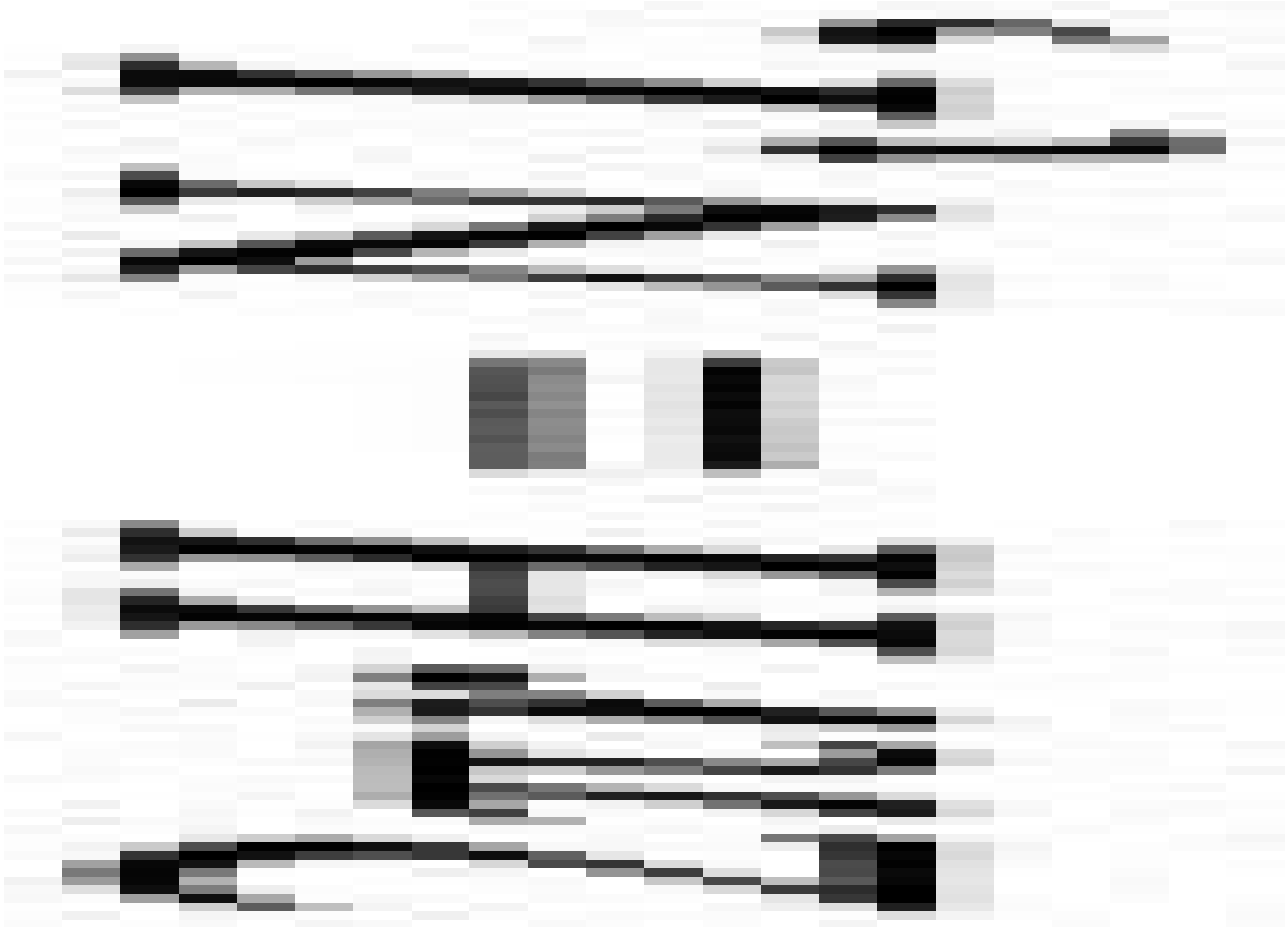
Consider a toroidal sample of magnetic material overwound uniformly with closed-packed coils of wire that carry a current I, as shown in Figure 5.13. Take a circular path of radius r inside the toroid. The line integral of the magnetic intensity around this path in a clockwise direction is equal to NtI where Nt is the total number of turns on the toroid,

Figure 5.13. A toroidal shaped magnetic material, uniformly overwound with coils carrying a current.

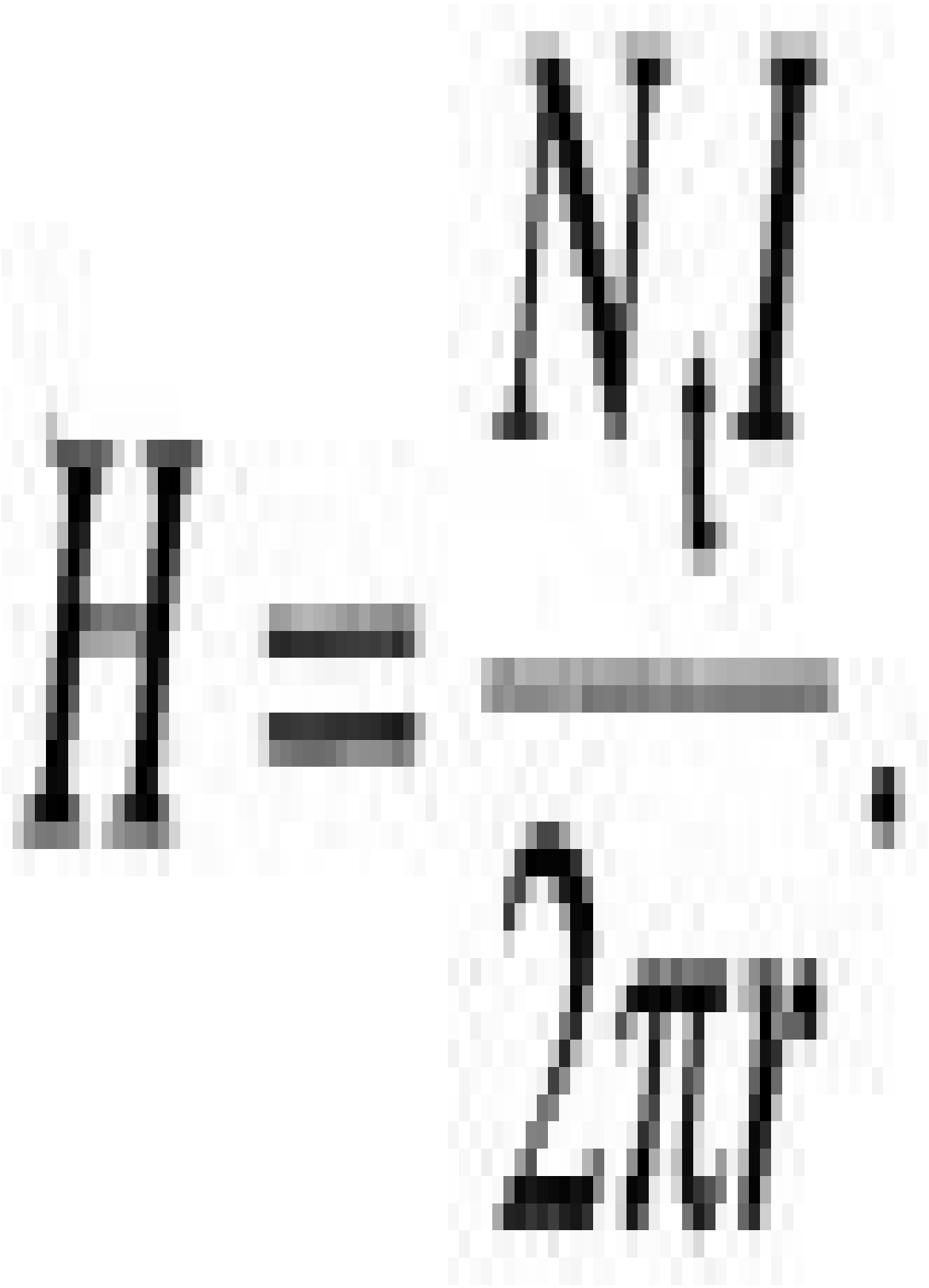

r'
r
b
a
H
I

(5.40)

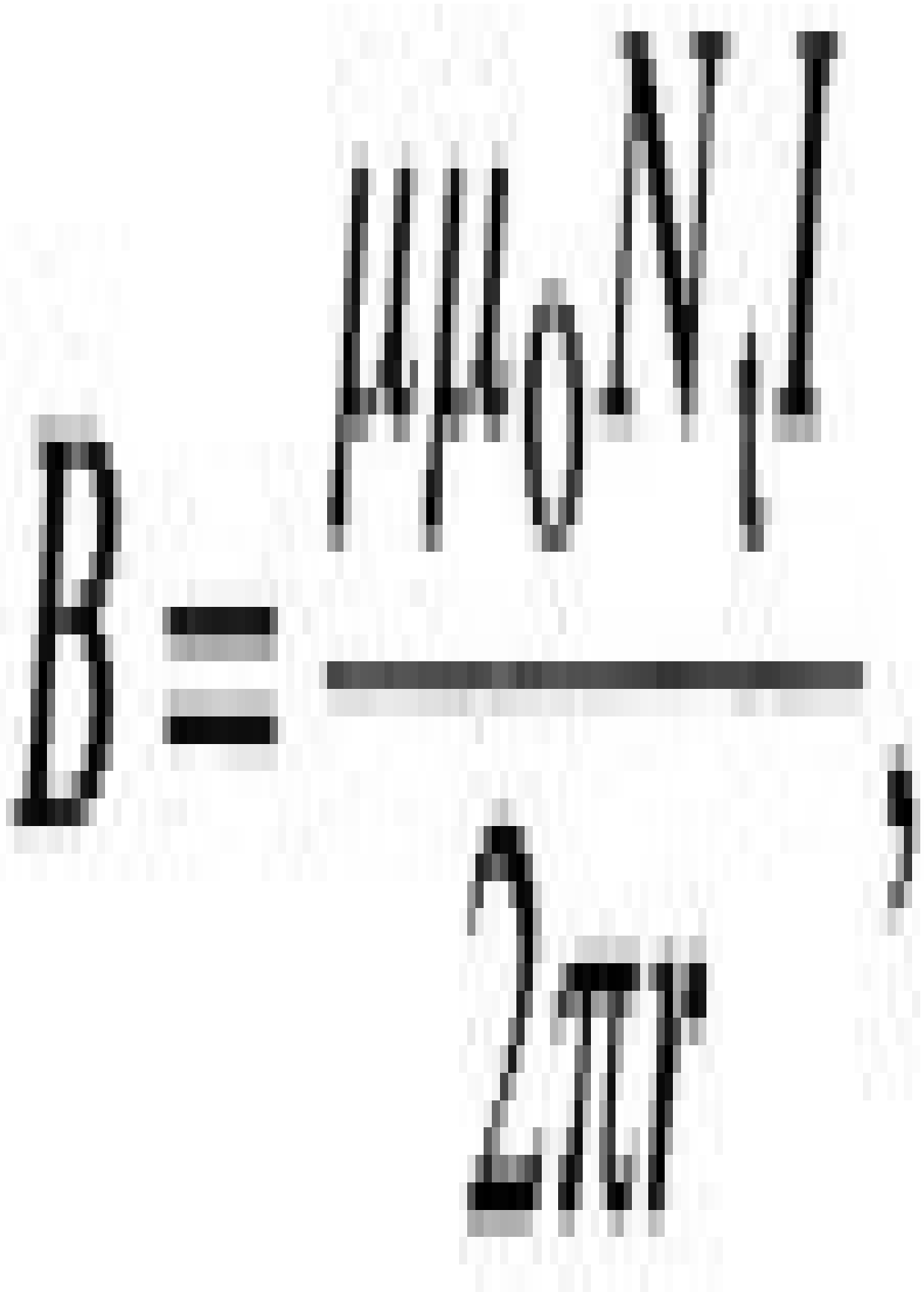
But by symmetry the magnetic intensity must have the same magnitude at all points on this circle and be tangential to the circle, hence



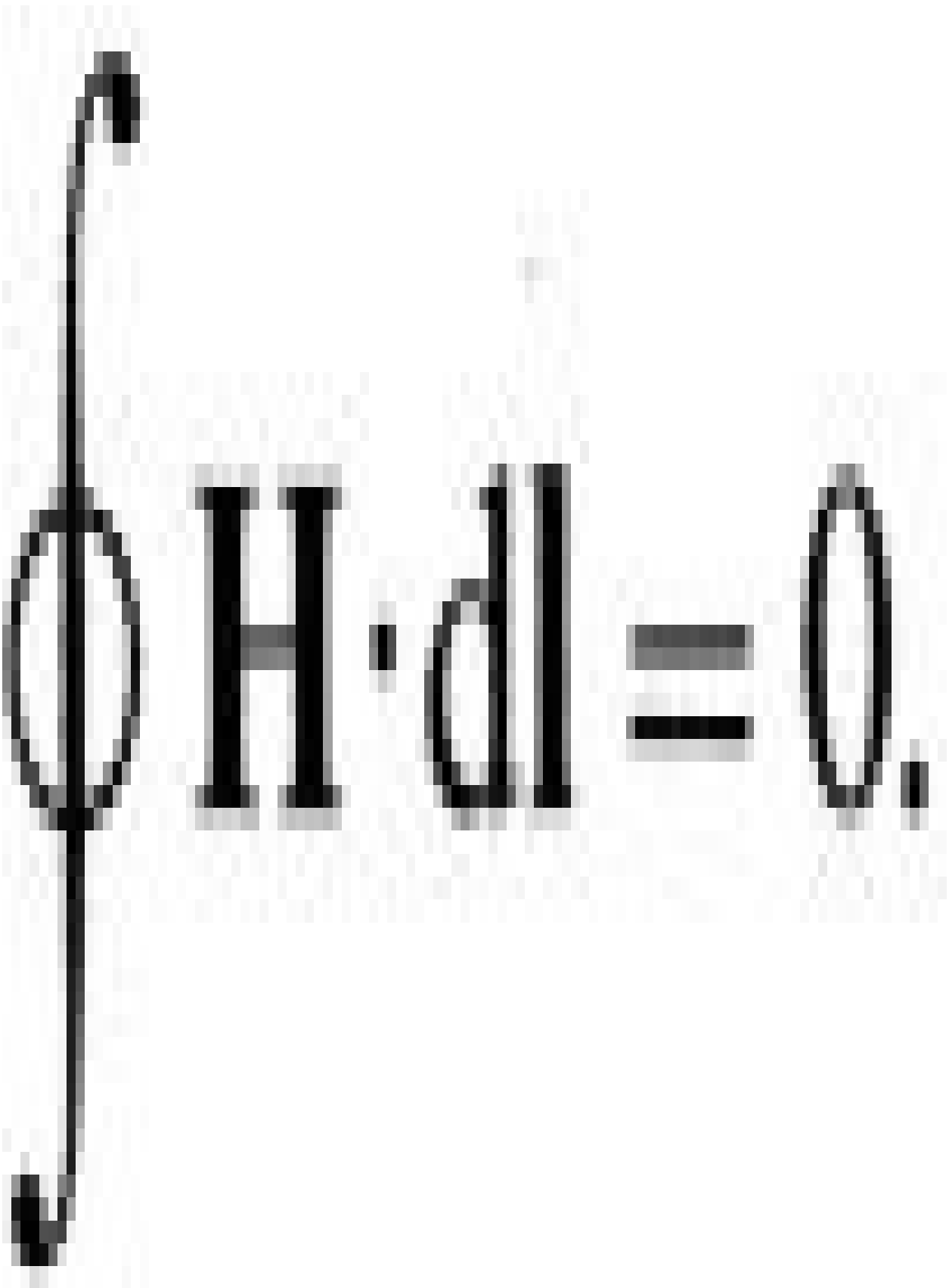
and



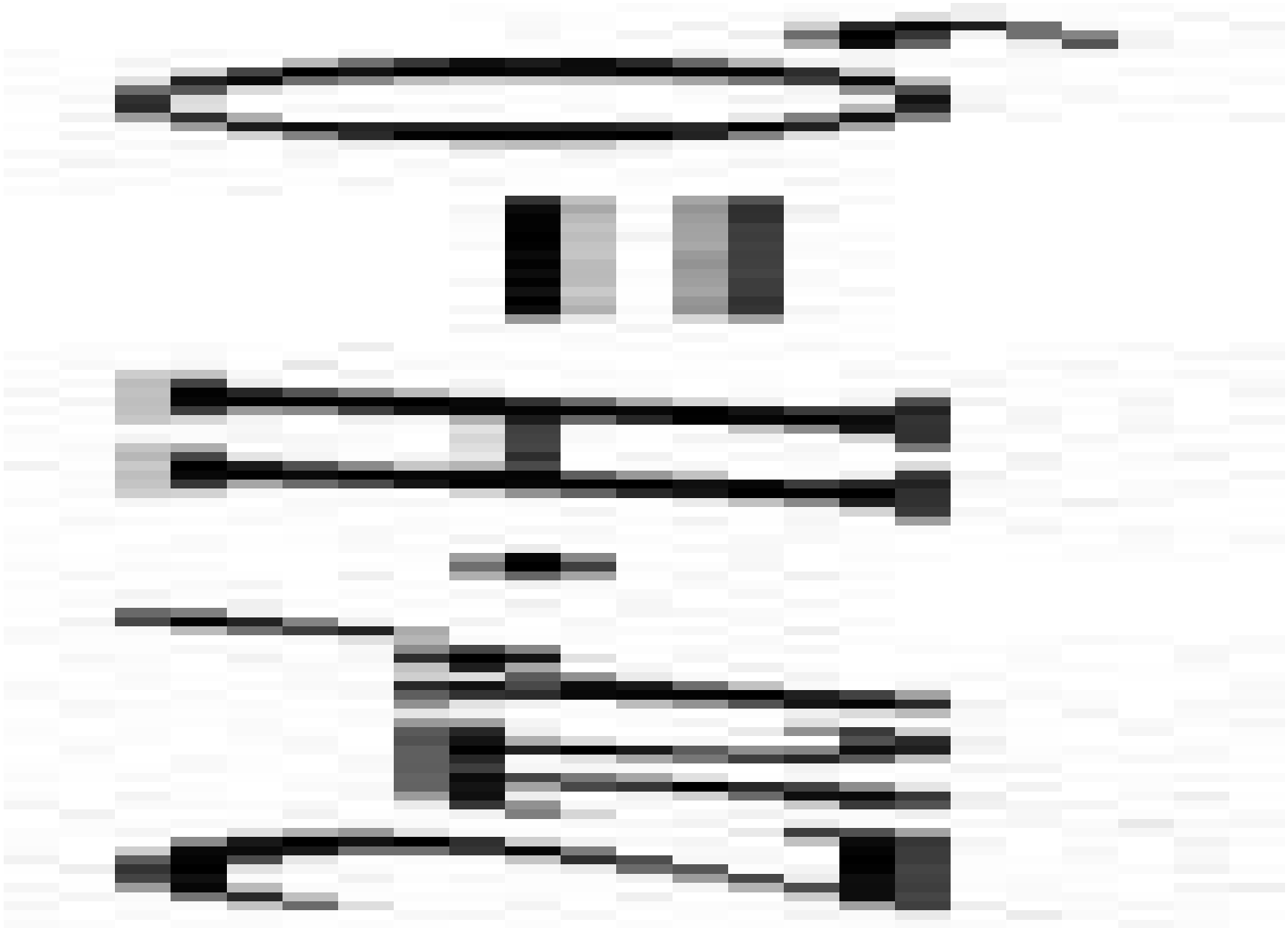
The magnetic field B inside the toroid is thus given by



where μ is the relative permeability of the material. Consider a different circular path of radius r' outside the coils as shown in Figure 5.13. If there were a field H outside it would again be tangential and have the same magnitude at all points on the circle. But now the total current which passes through the circle is zero and so from Equation (5.40)



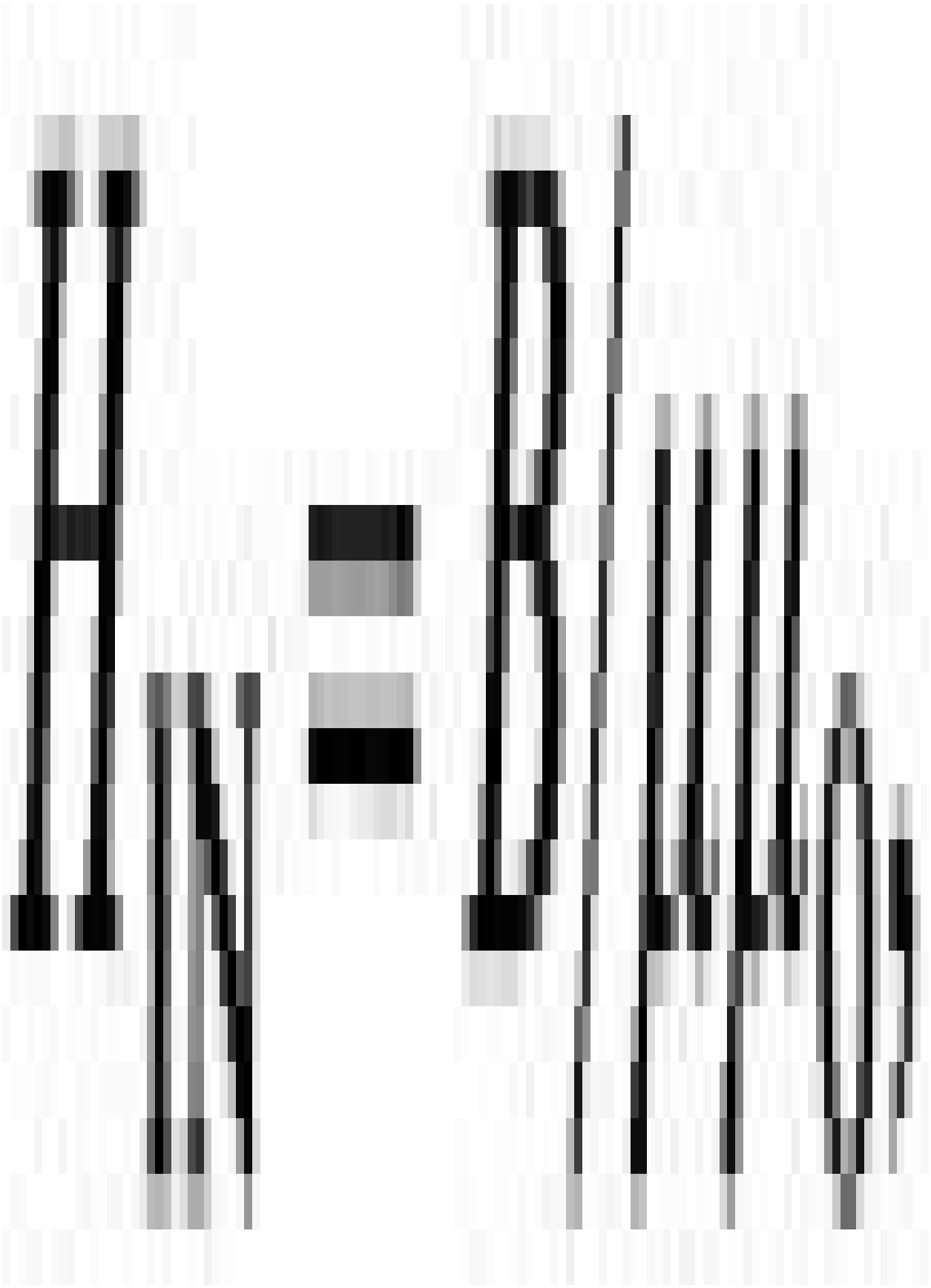
Hence



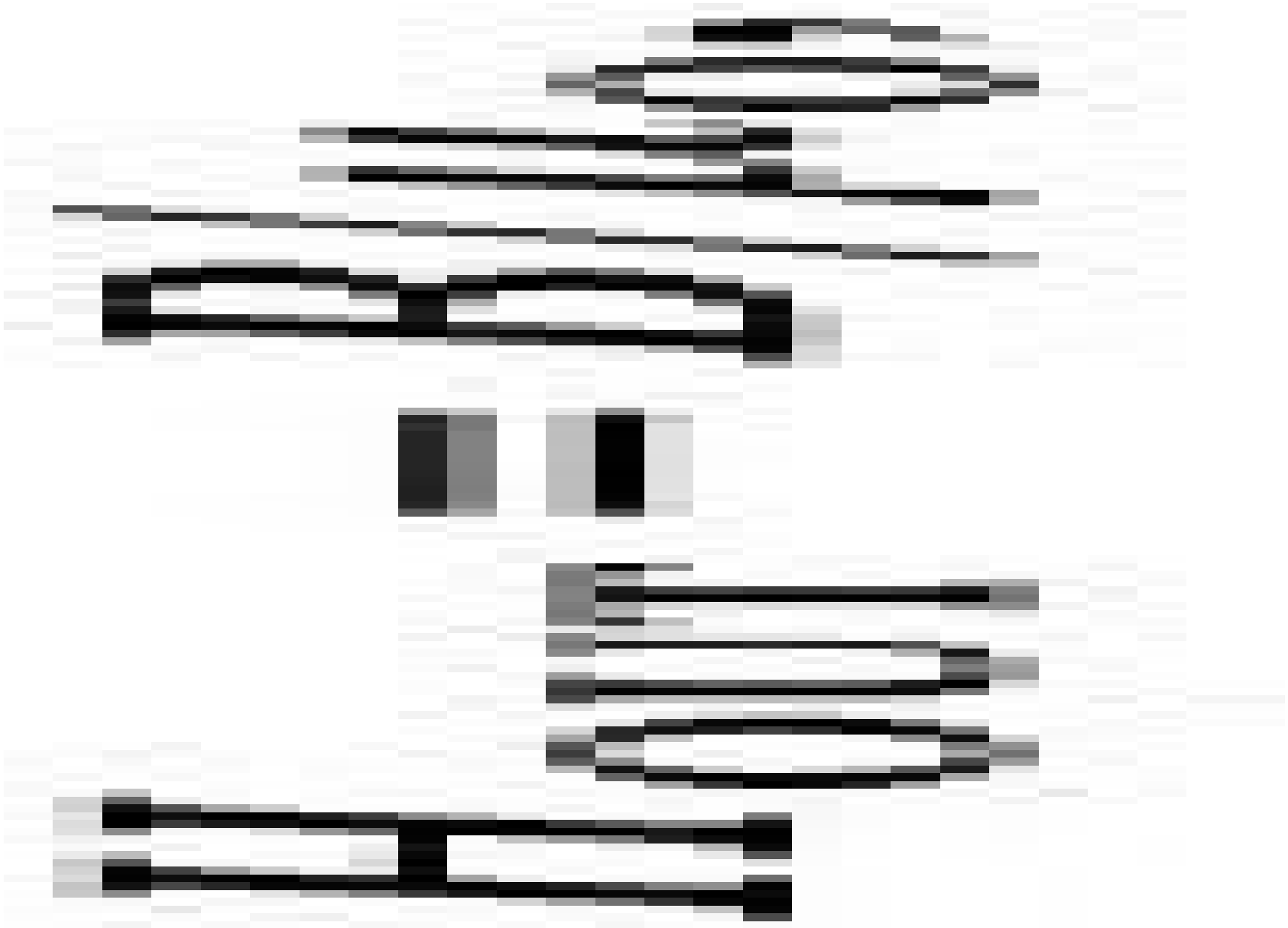
and the field H outside is zero. The same argument can be applied to any point outside the coils, so that outside, the fields are everywhere zero.

If the sample were removed from within the coils Equation (5.40) would still be true. To a good approximation the field H would again be tangential to circles centred on the middle. The magnetic intensity would thus be unchanged, but the magnetic field B would be reduced by the factor μ.

Suppose now that there is a small gap of length I cut in the toroidal sample, as shown in Figure 5.14. An arrangement similar to this is often used to make an electromagnet. If the core is made of a ferromagnetic material the field B in the gap is greatly enhanced over its value in the absence of the core. How do we determine the field in the gap? As a first approximation we will assume that there are still no fields outside the toroidal region of space occupied by the core and gap, i.e. that all the magnetic flux produced by the current in the coils and the magnetization of the material remains inside the coils. If this is so, then again the fields are tangential to a circle of radius r, as drawn, and at all points on this circle within the material the magnetic intensity has the same magnitude. At the interfaces between the material and the air in the gap the magnetic fields are normal to the boundary, and so the magnetic field B has the same magnitude B in the medium and in the gap. If the relative permeability of the material of the sample is μ, the magnetic intensity inside the sample is

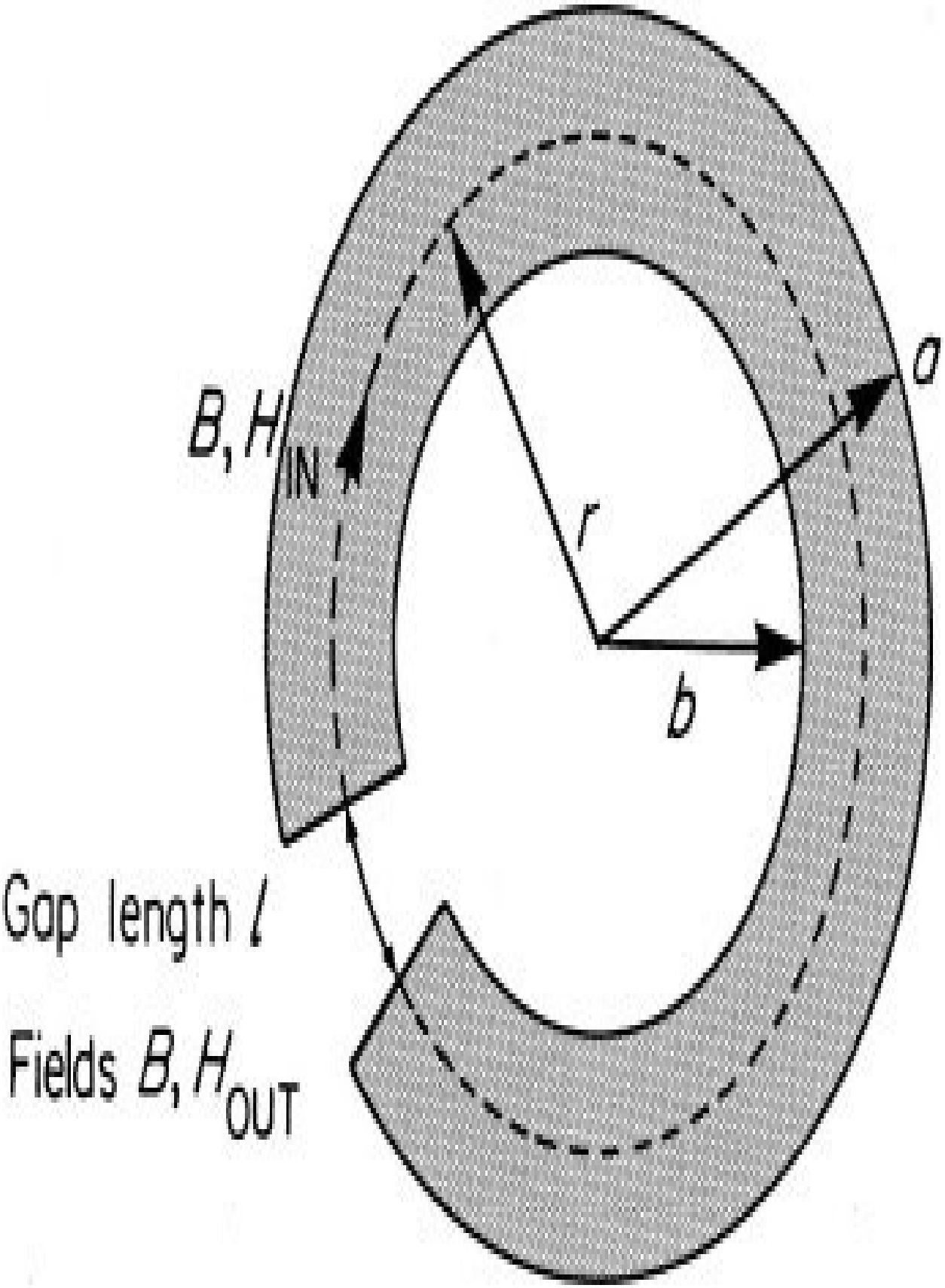


and the magnetic intensity in the gap is

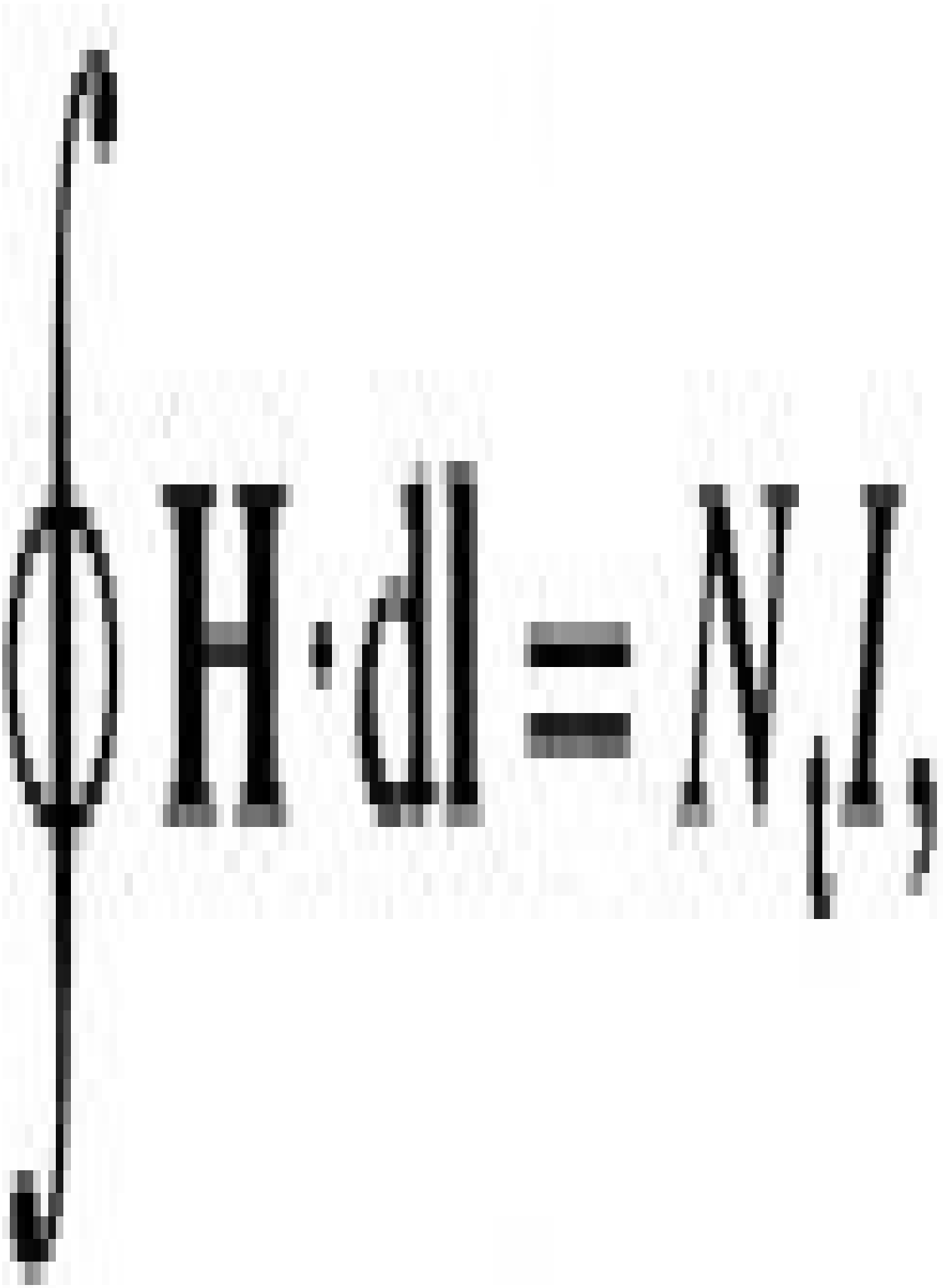


if we put the relative permeability of air equal to unity. Taking the complete path

Figure 5.14. A toroidal shaped magnetic material with a wedge-shaped gap cut in it. The toroid is uniformly overwound with coils carrying a current, as in Figure 5.13.


B, H
IN
a
r
b
Gap length l
Fields B, H
OUT

around the circle



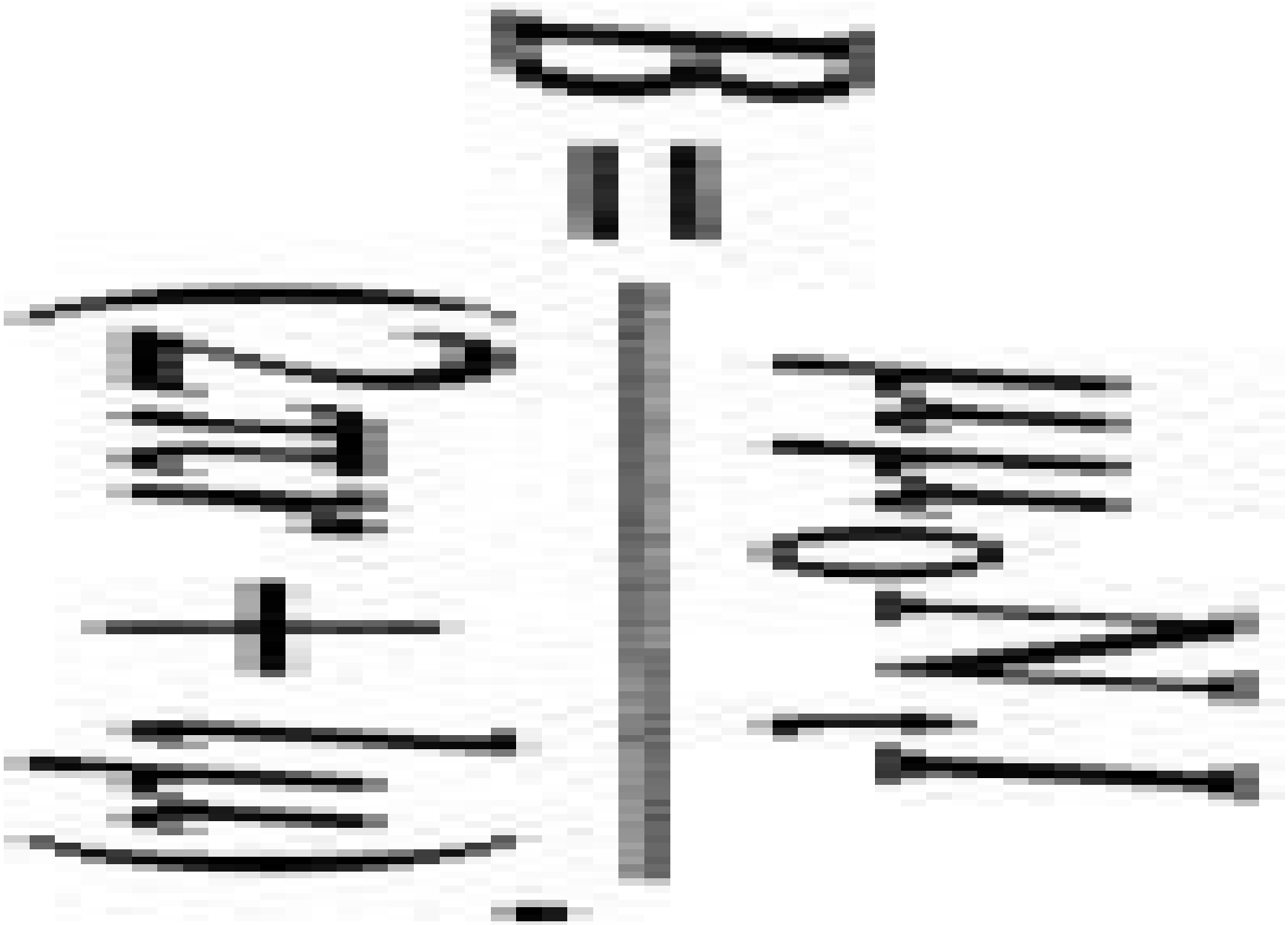
where Nt is the number of turns on the winding and I is the current. Evaluating this line integral gives

(5.41)

or

(5.42)

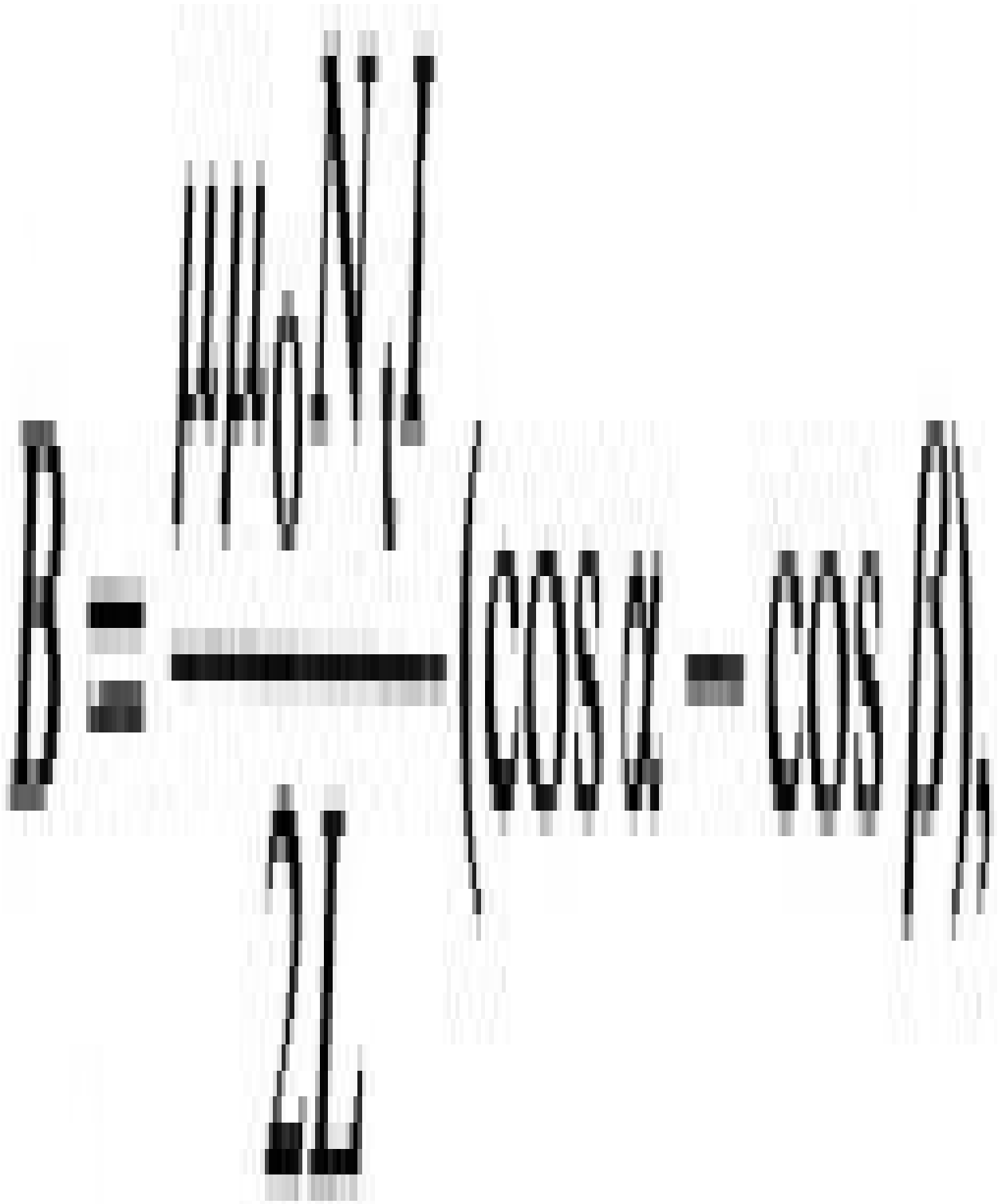
For ordinary materials the relative permeability μ is very near unity and Equation (5.42) reduces to the answer we obtained for a complete toroid. For ferromagnetic materials μ depends on the fields H and B. However, even though μ is not constant, we may still write B = μμ0H and the equations above still apply. For fields B less than about one tesla the value of μ, defined by the equation B = μμ0H is typically in the region of 1000. The field is then approximately given by



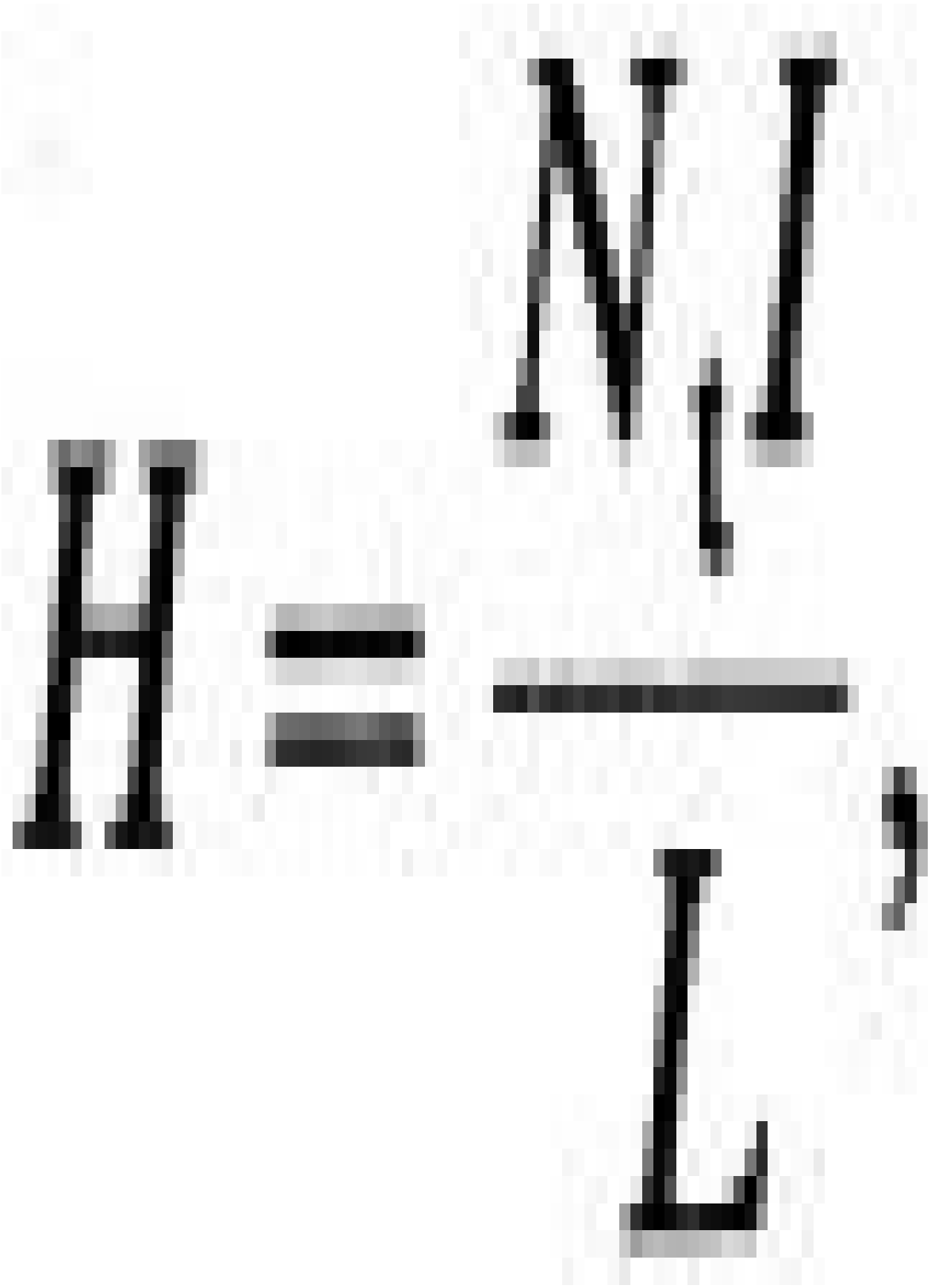
Unless the gap is very small, lμ » 2πr and

As we expect, a ferromagnetic material, due to its large magnetization, increases the field B in the gap by a large amount in comparison with the value $\mu_0 NtI/2\pi r$ in an empty coil.

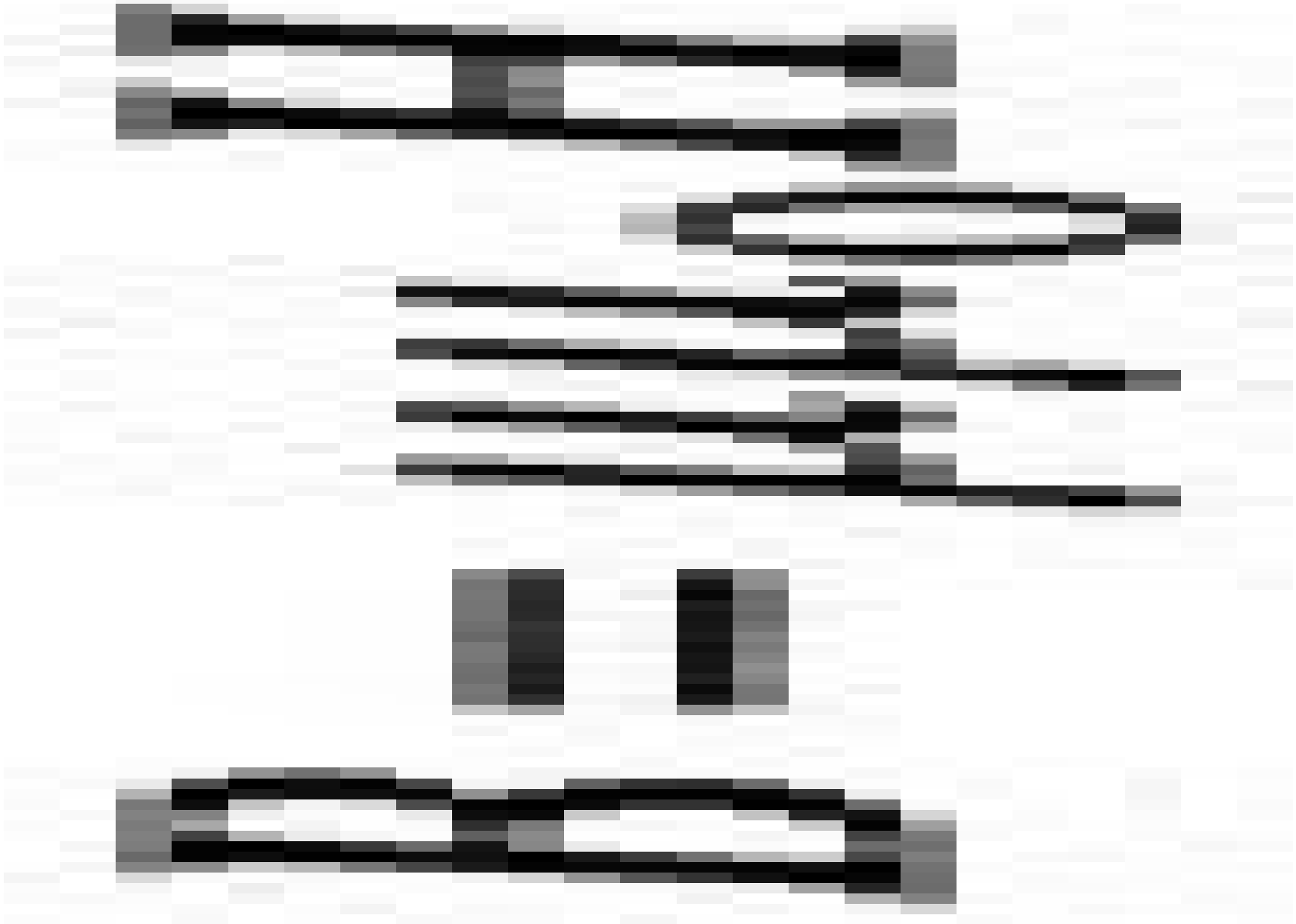
In order to calculate more precisely the field in the gap of a particular electromagnet it is necessary to know the relation between the fields B and H within the material, i.e. to know the value of the permeability μ at each value of the field H. The relationship between B and H is given by the magnetization curve, often called the B–H curve, of the material. The B–H curve for soft iron, a material often used in electromagnets, is shown in Figure 5.15. This curve can be obtained approximately by measuring the field B near the end of a rod of soft iron which fills a long solenoid. The field at a point on axis near the end of the solenoid is given by (see Problem (5.8))



where the angles α and β specify the position of the point, Nt is the total number of turns and I the current. L is the length of the solenoid. The magnetic intensity inside the rod can be determined to be roughly



by using Equation (5.38) and making the usual assumptions of neglecting the fields outside the solenoid and taking them to be uniform inside. The field B measured near the end of the rod is approximately equal to that inside the rod because of the continuity of the normal component. Thus if the field B is measured as a function of the current through the solenoid, I, the permeability of the rod can be determined at each value of the magnetic intensity. Since



the B–H curve for the material can be plotted. Another method, similar in principle but more accurate, would be to construct a toroidal electromagnet such as shown in Figure 5.14 out of the material. The B–H curve can be determined by relating the field B measured in the gap (which is the same as the field in the material) to the field H in the material deduced from the current I using Ampère's integral law for the field H.

If the electromagnet of Figure 5.14 were made with soft iron, the field in the gap could be calculated with the aid of the B–H curve for soft iron shown in

Figure 5.15. The relation between the magnetic field B and the magnetic intensity for soft iron.

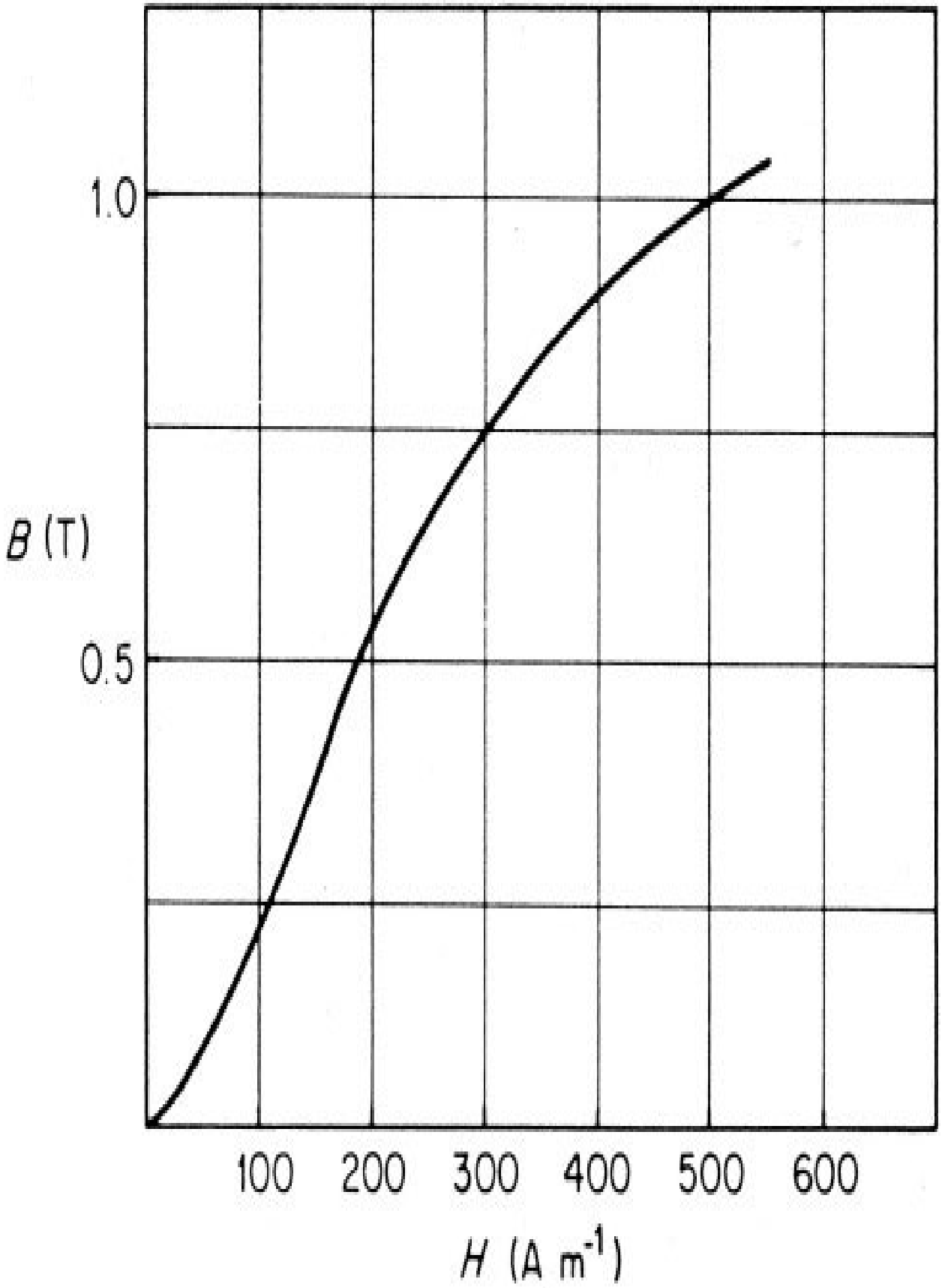
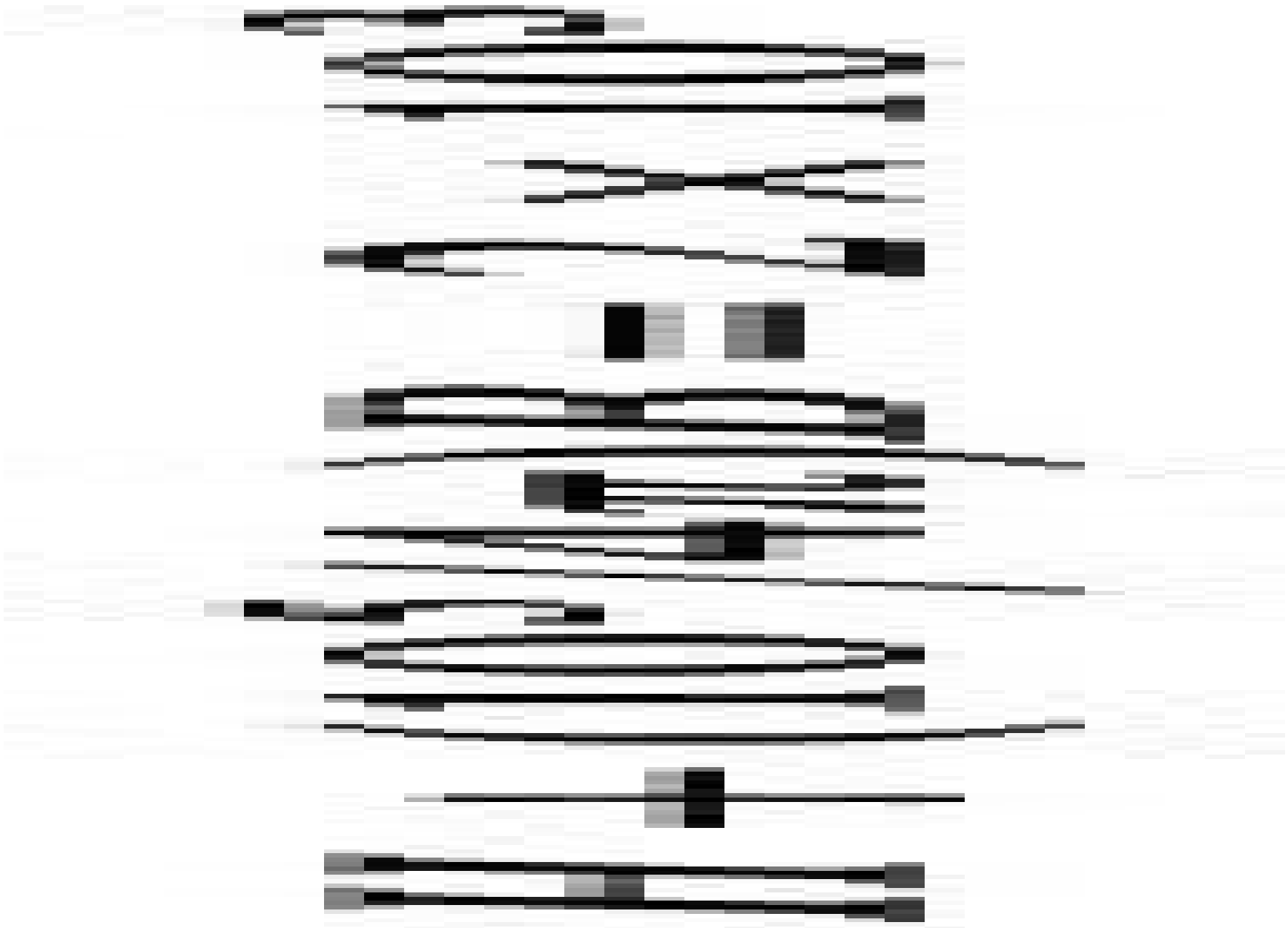

1.0
0.5
B (T)
100
200
300
400
500
600
H (A m-1)

Figure 5.15. Rewriting Equation (5.41) in the form

(5.43)

we have a linear equation which the intensity H and the magnetic field B within the material have to satisfy. Figure 5.15 shows another relation which the fields have to satisfy, and their actual values inside the material can be obtained graphically by finding the intersection of the straight line (5.43) with the B–H curve. For example if the product of the number of turns on the solenoid, Nt, and the current, I, is 2000 ampere turns; the gap length, l, is 1 cm, and the circumference of the toroid less gap is 1 m, the magnetic field B in the gap is given by the intersection of the line



with the B–H curve. The field is about 0.24 T, as shown in Figure 5.16.

Real electromagnets are not always made with the material in the form of a toroid, or with coils wound all around. A realistic, but still simple, design is one in which the yoke, or core, of the magnet is C-shaped, as in Figure 5.17. Large magnets are often H-shaped, as in Figure 5.22. The field in the magnet gap can be calculated roughly by making the same assumptions as before, viz. that all of the magnetic flux lies inside the iron and that the lines of the field B follow the contours of the yoke. The total magnetic flux Φ through any area obtained by slicing through the iron core of the magnet shown in Figure 5.14, for example, is then constant and given by

Figure 5.16. The calculation of the magnetic field in the gap of a soft iron electromagnet.

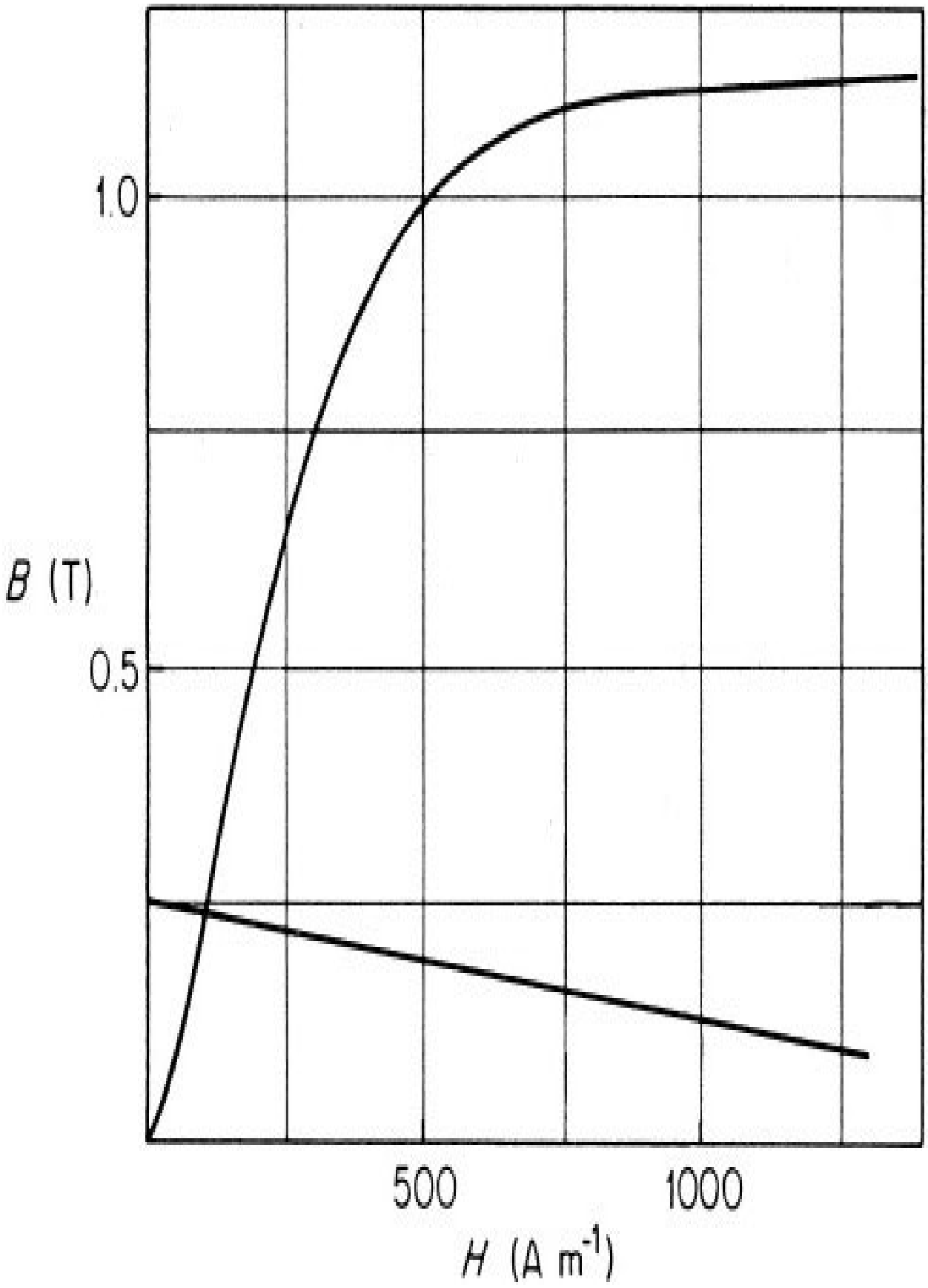
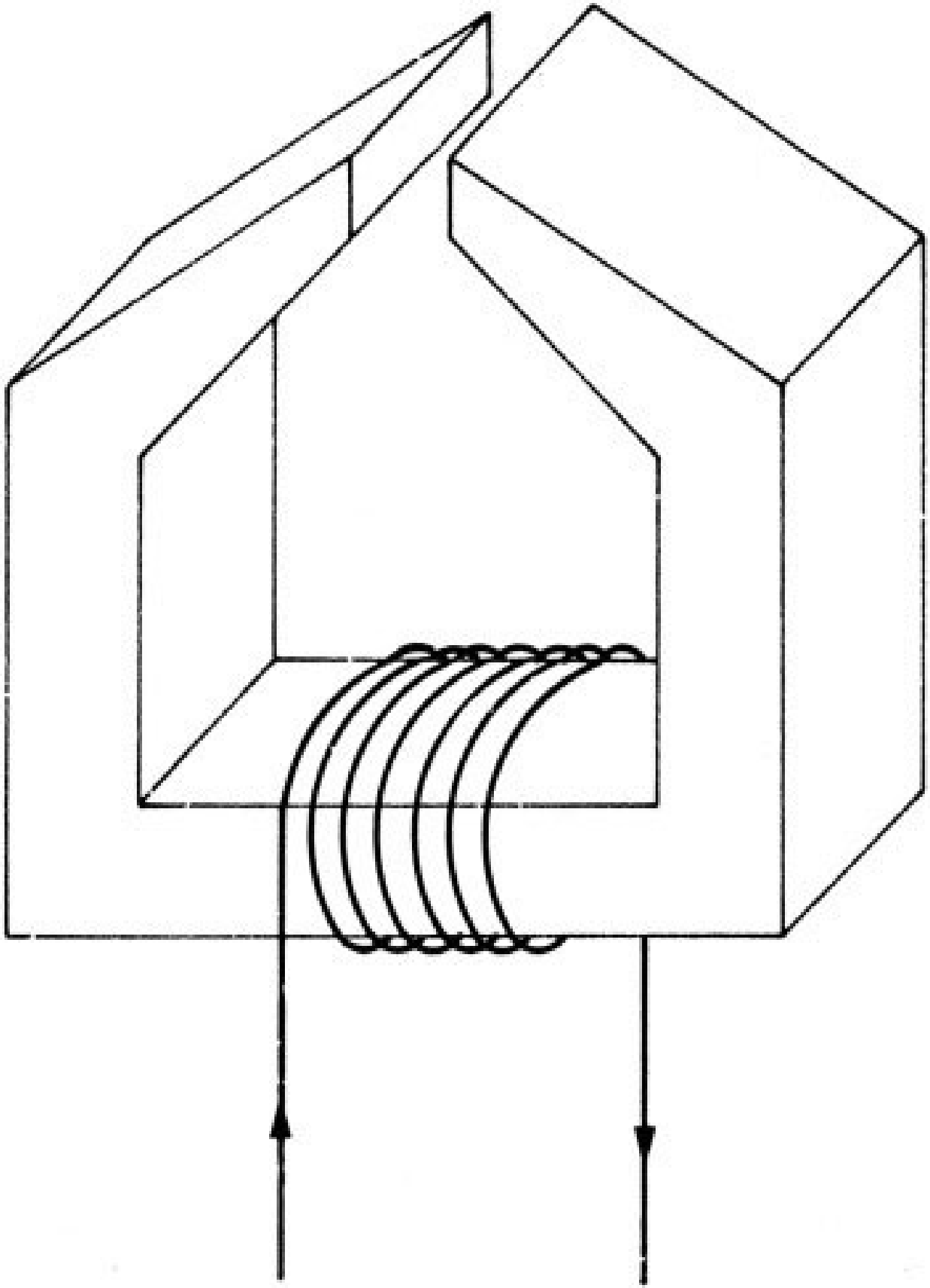
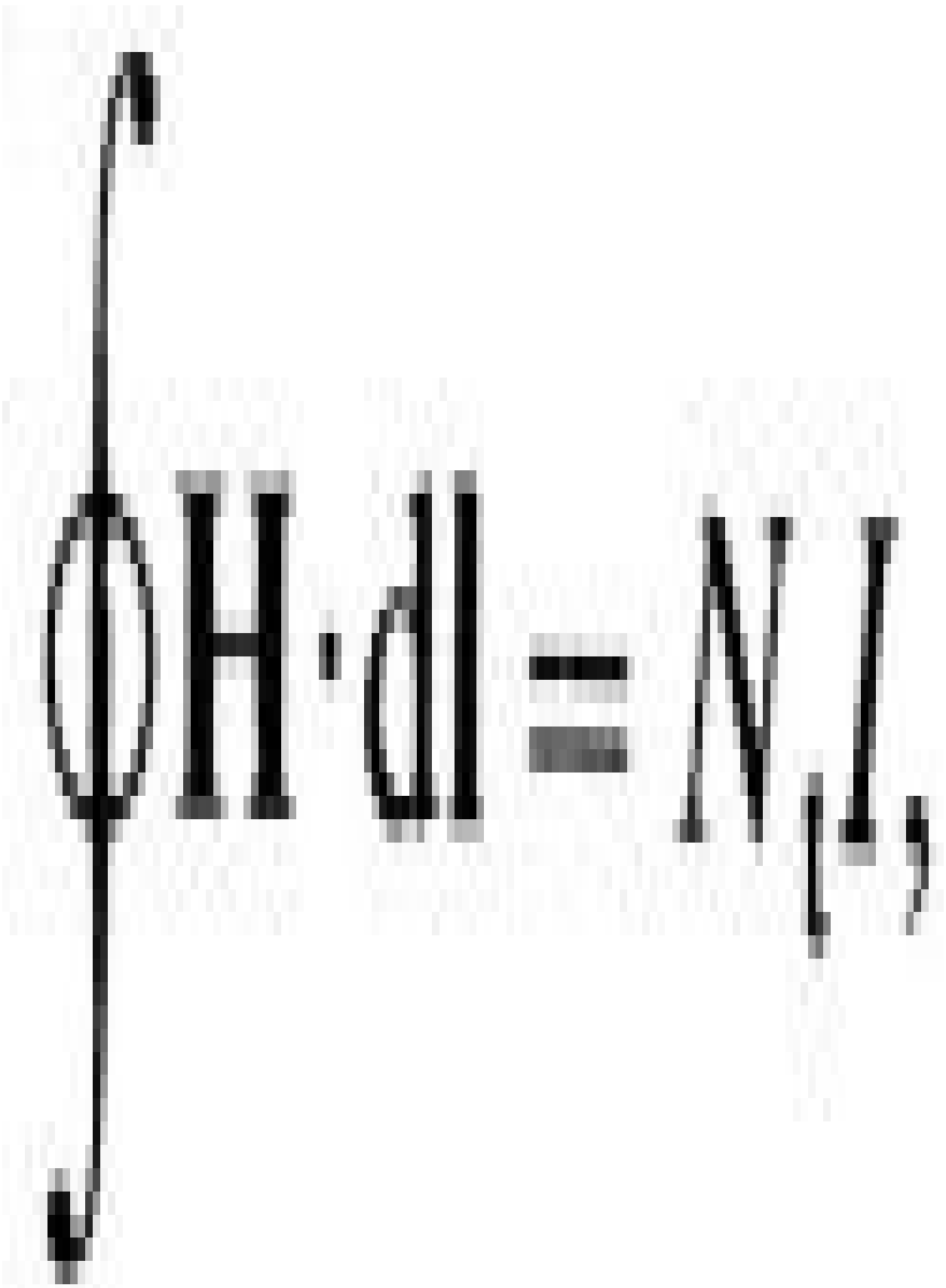

1.0
0.5
B (T)
500
1000
H (A m-1)

Figure 5.17. A simple electromagnet such as might be found in a tape recorder. The dimensions of the magnet in that case would typically enable it to fit easily into a cube of side 1 cm. The field in the gap would be less than that given by Equation (5.45) due to flux leakage around the pole faces.

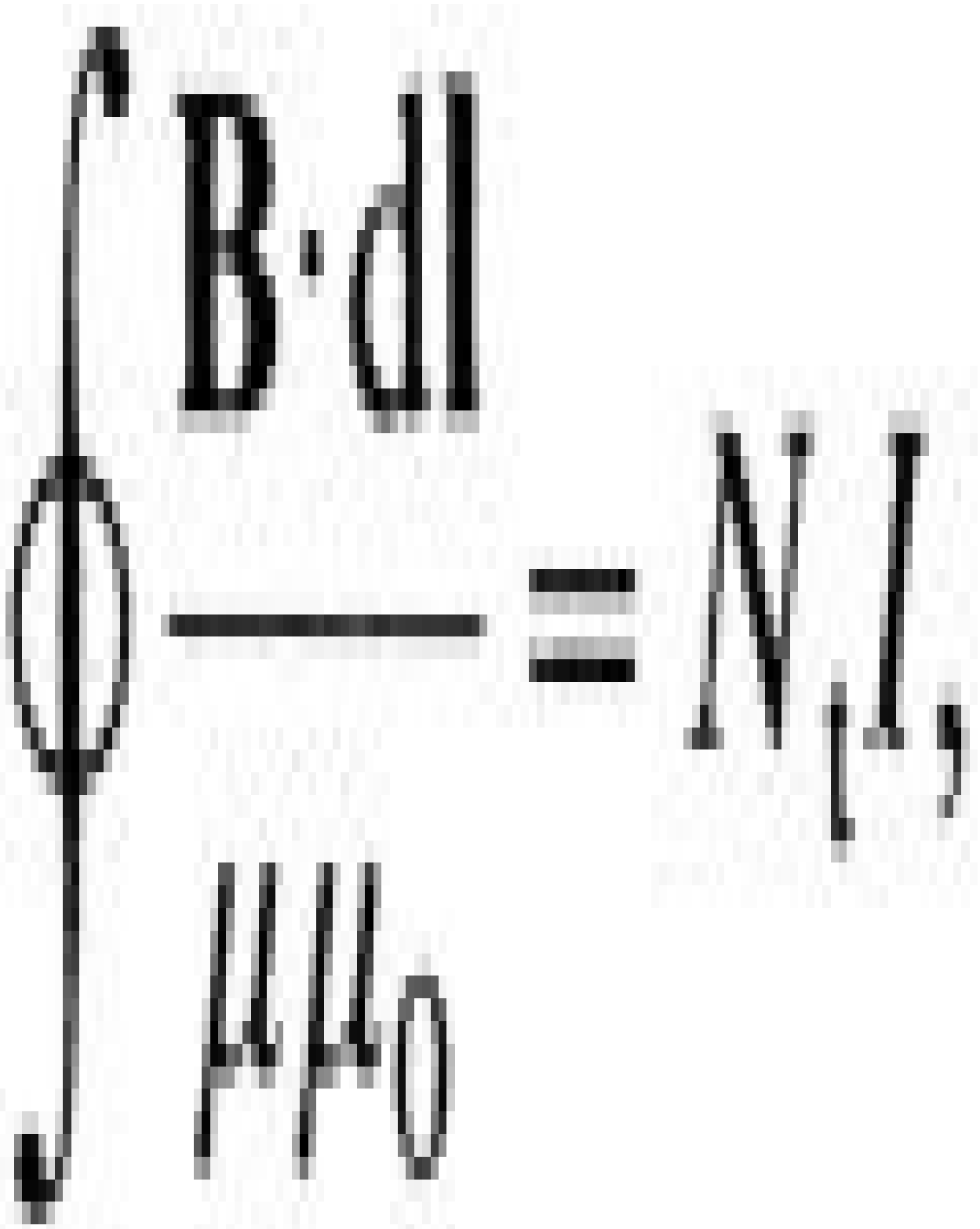


(5.44)

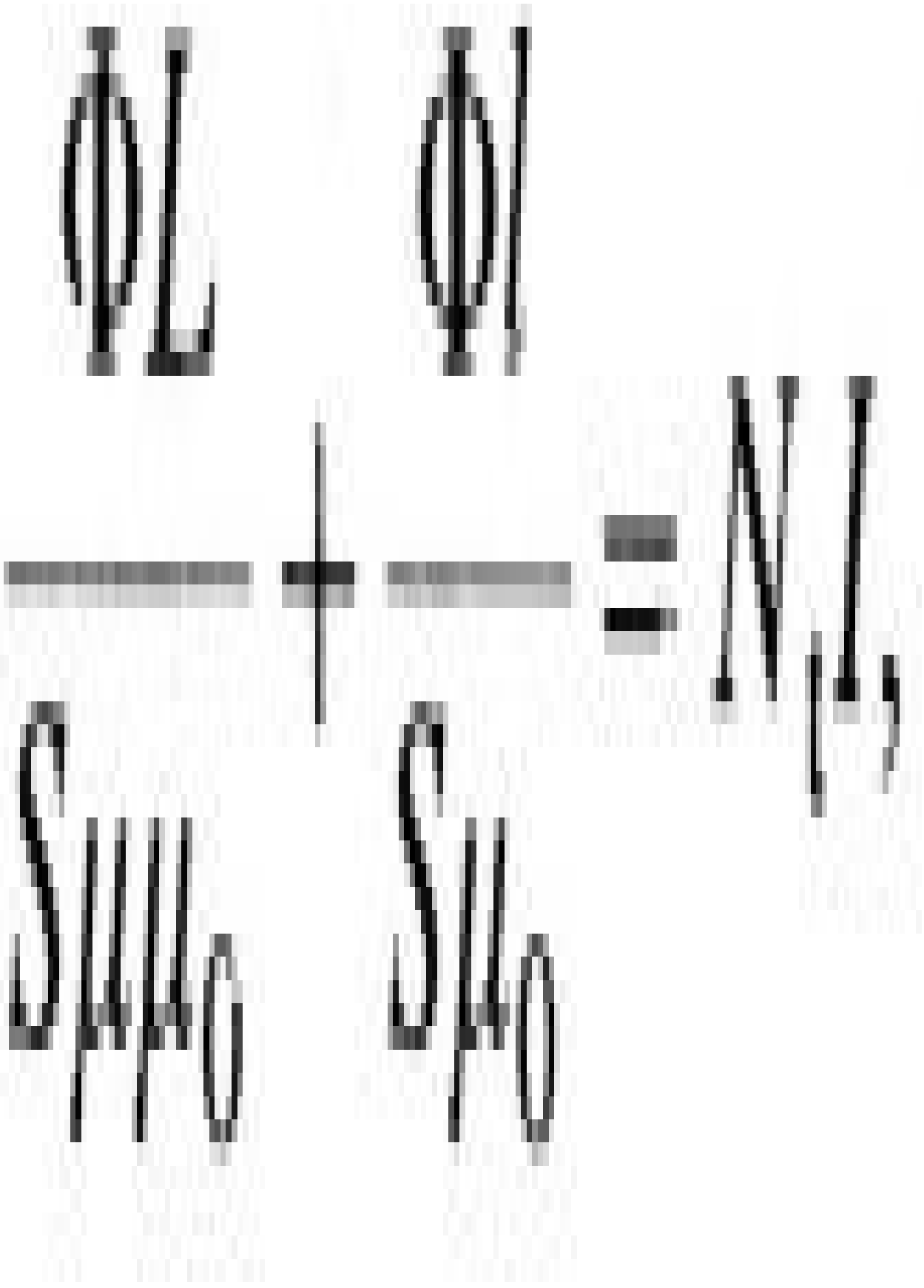
where B is the field at the place where the cross-sectional area of the yoke is S. For a complete path drawn through the yoke and gap,



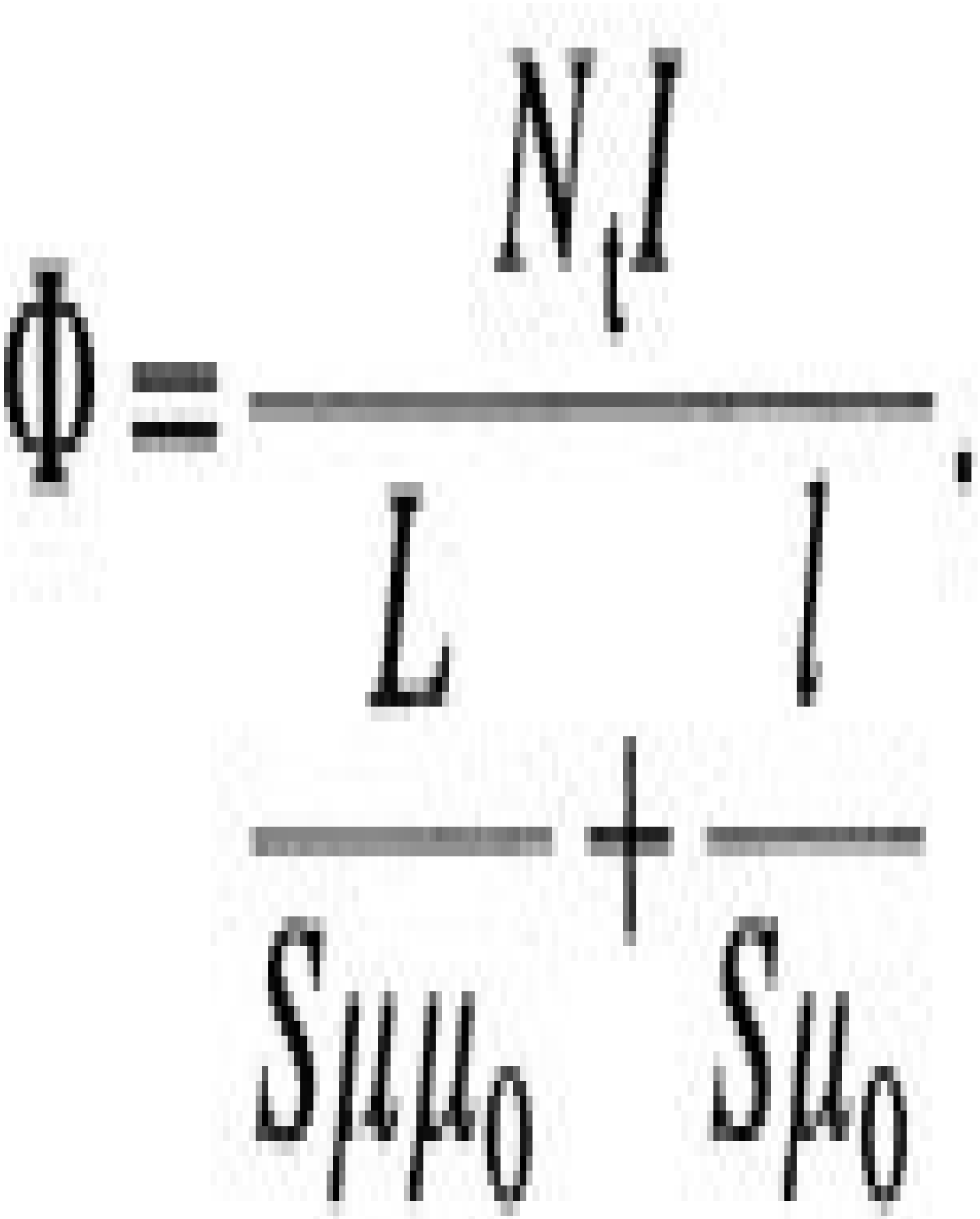
where the symbols have their usual meanings. Rewriting the line integral in terms of the field B we have


∮ B·dl
= N_I
μ0

or



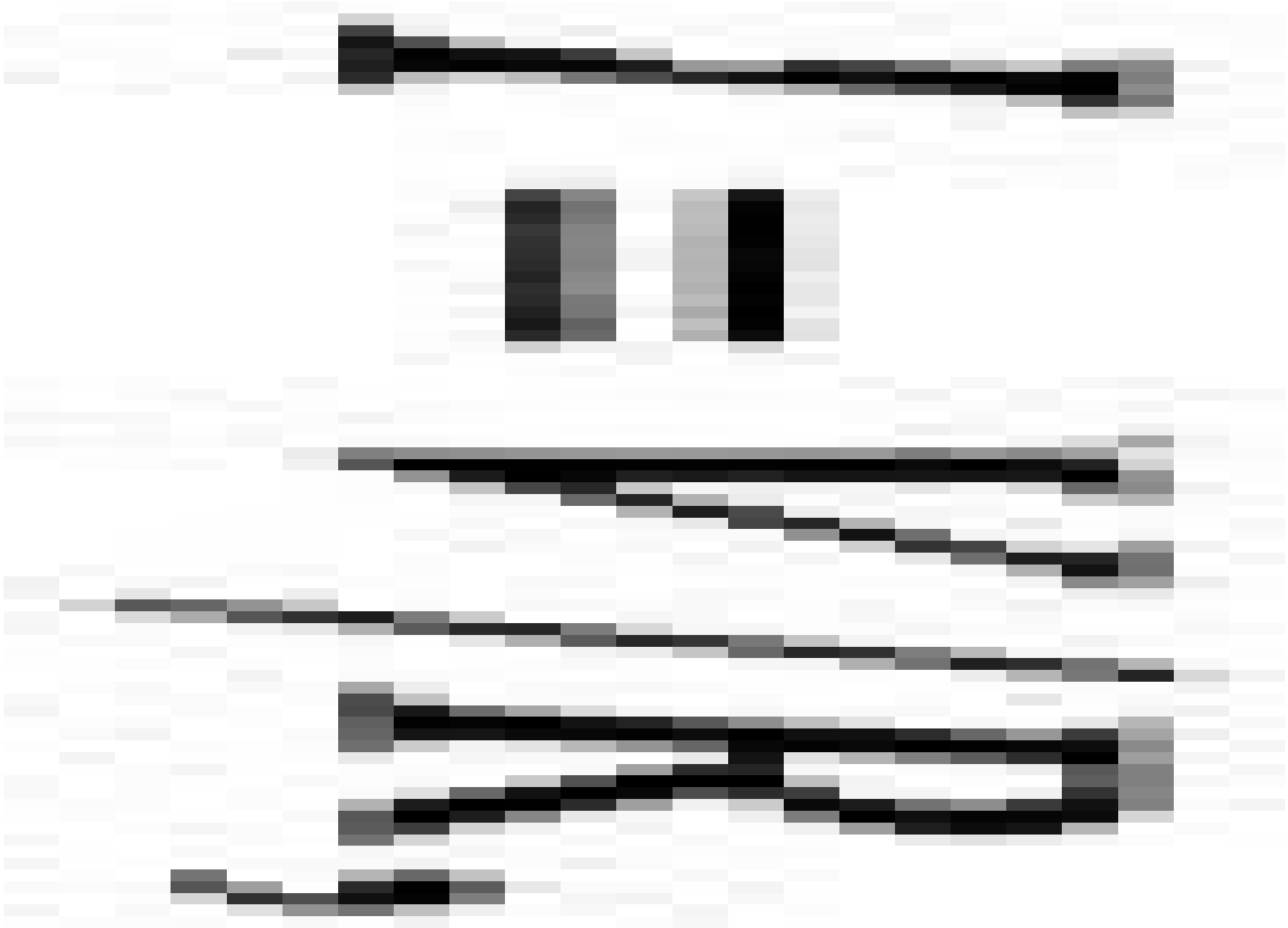
from Equation (5.44), where L and l are the path lengths in iron and gap respectively. Hence



More generally if a magnet has n segments of lengths ln which have different cross-sectional areas Sn and/or relative permeabilities μn, using the same assumptions as above it can easily be shown that the flux Φ is given by

(5.45)

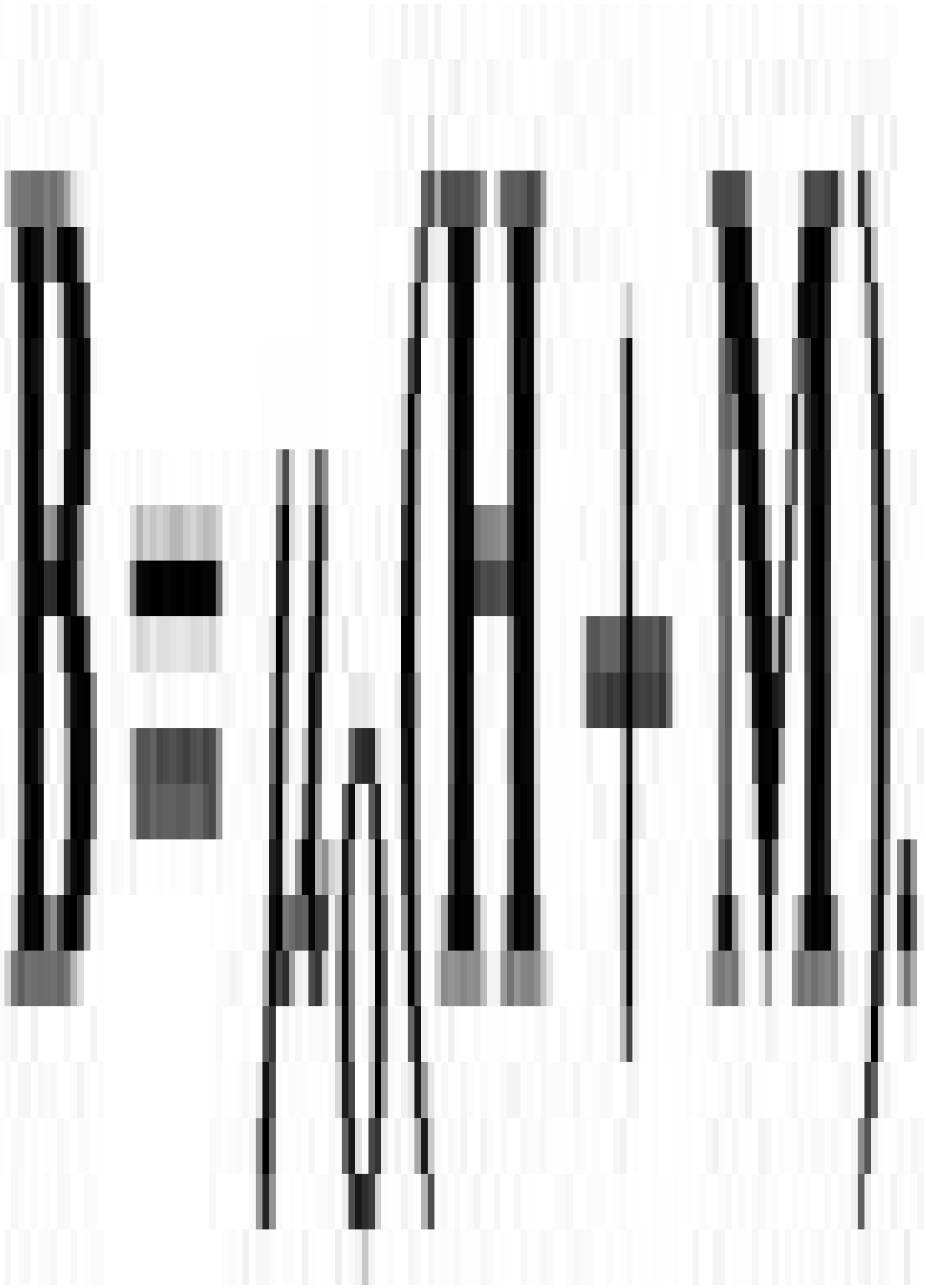
In Equation (5.45) the quantity (NtI)is often called the magnetomotive force (m.m.f.) in the magnetic circuit, by analogy with the e.m.f. V provided by a battery in an electrical circuit. Equation (5.45) is very similar to Ohm’s law



which gives the current I through a series of resistors Rn in series, which have a total resistance

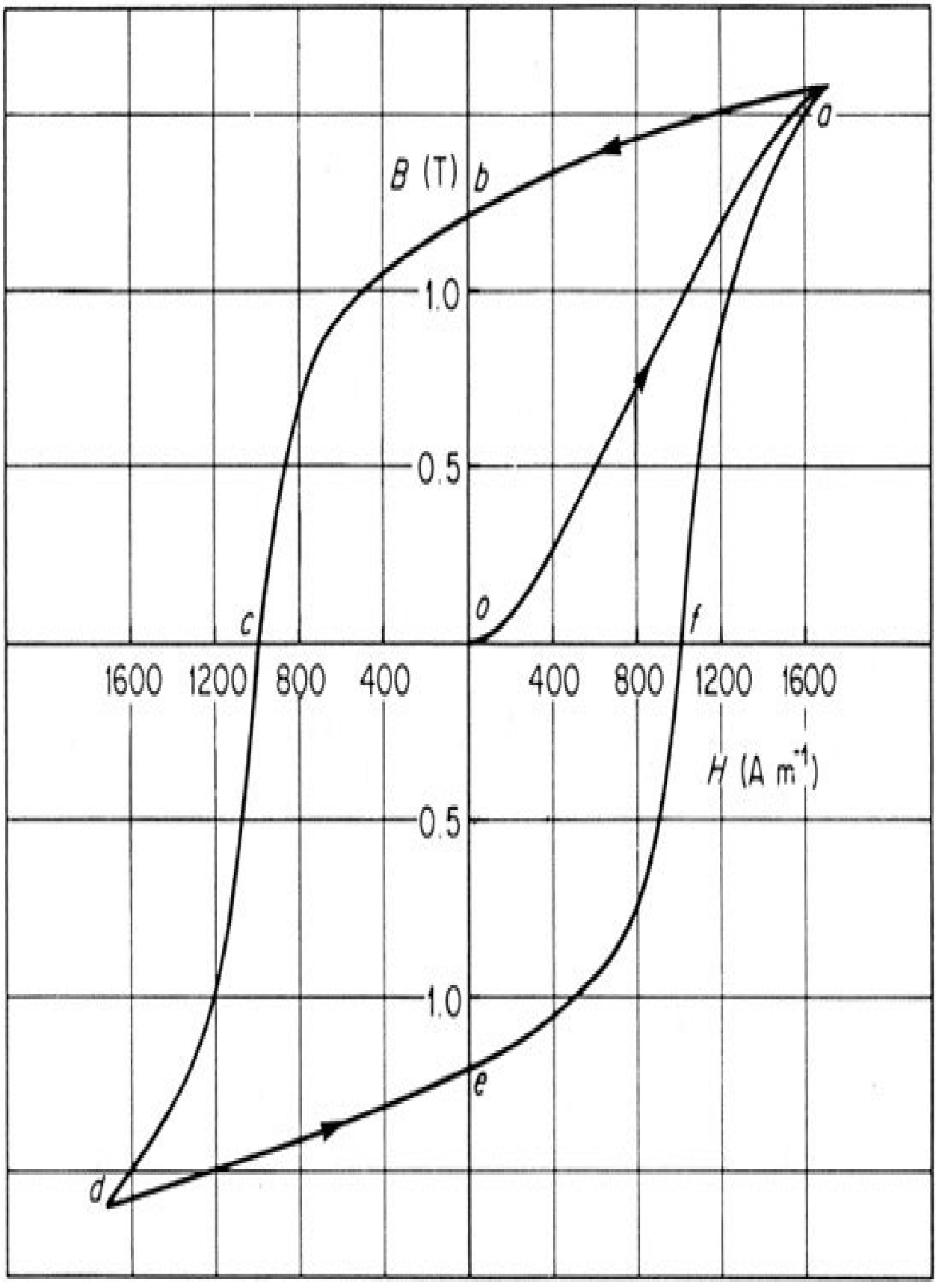
The quantities ln/μnμ0Sn are called the reluctances of the various segments.

Soft iron is used in electromagnets because when the current is switched off the fields in the gap and the iron become very nearly zero. If steel is used there remains a field even after the current has been reduced to zero. The B–H curves for steels and some other materials, which are called hard ferromagnetics, look like Figure 5.18. The numbers on the figure are for a typical hard steel. Suppose such a B–H curve is measured by putting a steel rod inside a solenoid, as discussed earlier. As the current is increased from zero, the curve follows the line oa. It levels off at a when the steel becomes saturated. At that point the magnetization M of the steel has reached its maximum value, its saturation magnetization Ms, when all the individual moments are aligned. The magnetic field B increases after the point a only by virtue of the term μ0 H in the equation



If the current is reduced below the saturation level and back to zero, the B–H curve follows the line ab. For currents in the reverse direction the magnetic intensity H needs to be quite large for the field B to be reduced to zero at the point c. At the point d we have reached saturation in the reverse direction. The whole curve abcdefa is called the hysteresis curve of the material. The value of the field B at the point b on the curve is called the remanence of the material: if this is large the material can be used in a permanent magnet, which produces a flux of magnetic induction even in the absence of external currents. For soft iron the remanence is very nearly zero.

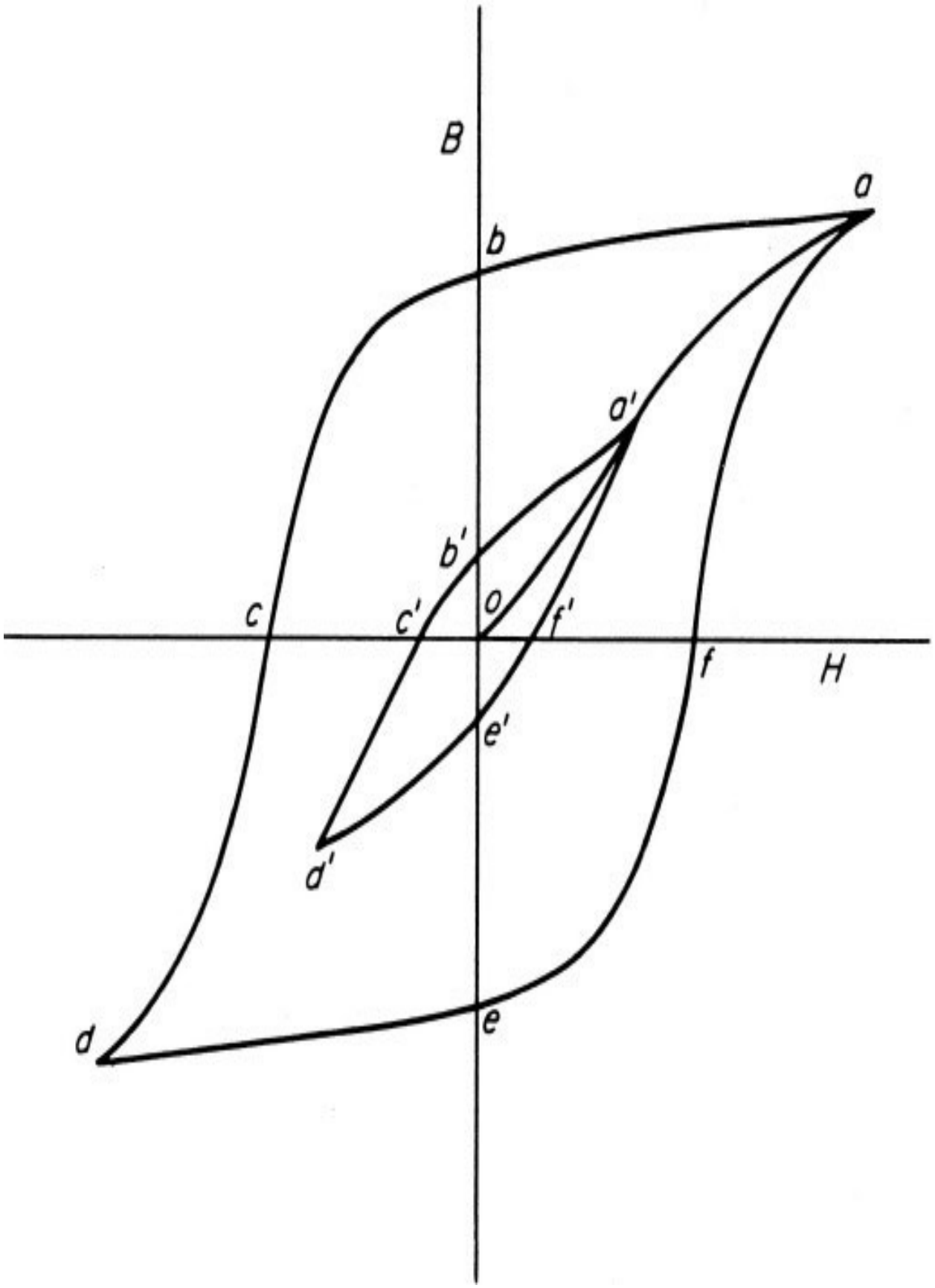
Figure 5.18. The B–H curve for a typical hard steel.


B (T)
1.0
0.5
0.5
1.0
a
b
c
d
e
f
1600
1200
800
400
400
800
1200
1600
H (A m⁻¹)

The hysteresis curve shown in Figure 5.18 is the major hysteresis curve for the steel, in which the steel is taken up to saturation for each direction of the field H. If the sample is not taken up to saturation along the initial portion oa it will return from the point a′ of Figure 5.19 via a path which in general has a different shape to the shape of the major hysteresis loop. There are an infinite number of minor hysteresis curves like a′ b′c′d′e′ f′a′ and so the magnetic field B in a hard ferromagnetic material can have many different values for the same value of the intensity H. The values of B and H that exist in a specimen depend on the particular magnetic treatment the specimen has undergone, i.e. on its magnetic history.

The hysteresis effect persists to some extent in the soft materials used in electromagnets and this has practical consequences. Suppose for example we wanted to measure the susceptibility of a paramagnetic material. This could be done by measuring the force on a sample placed in the gap of an electromagnet. If accurate measurements are required it is not sufficient to monitor only the current in the magnet. Identical currents will not always produce exactly the same magnetic field because of uncertainties due to hysteresis. The actual magnetic field in the gap must be measured.

Figure 5.19. The major hysteresis curve and one minor curve for hard steel.


B
a
b
a'
b'
c
c'
o
f'
f
H
e'
d'
d
e

The use of the approximations of the magnetic circuit for the calculations of fields due to electromagnets can produce errors of up to about 15%, or more if the iron is near saturation, when there is appreciable flux leakage out of the iron. The field lines in the gap bulge outwards, corresponding to a smaller field than given by the assumption that the field lines pass straight across the gap.

Saturation magnetization is approached for most ferromagnetic materials at fields near to 1.5 T. In order to obtain such fields, magnetic intensities approaching 10^4 amperes/m are required. These can be obtained for example by passing 20 A through a multilayered solenoid which has 10^3 turns/m, or 2 A through 10^4 turns/m.

Iron-cored electromagnets which produce large fields require large currents through many coils. These currents produce large power dissipations in the resistance of the coils resulting in large expenses for power losses and cooling. In addition the coils have to be of large cross-section to minimize resistance and hence heating losses and this makes the cost of the conductor, which would typically be of pure copper, very expensive. In recent years the technology of making coils out of certain superconducting materials has advanced considerably, and this has led to many large magnets, for example some used in particle accelerators or in other research fields, being superconducting magnets with coils which have exceedingly small resistance.

An example of a superconducting magnet, which illustrates their capabilities, is given by considering a solenoid made of wire containing superconducting material. At the present time the superconducting parts of the wire are usually made of a niobium-titanium alloy. When cooled to 4.2 K, the boiling point of liquid helium, the alloy becomes superconducting and has effectively zero resistance as long as current densities within it remain below about 4×10^{10} Am^{-2} and the magnetic fields at their surfaces stay below about 22 T. The actual magnet coils are made from thin copper wires in which are embedded filaments of niobium–titanium alloy. Each composite wire can pass currents of up to about

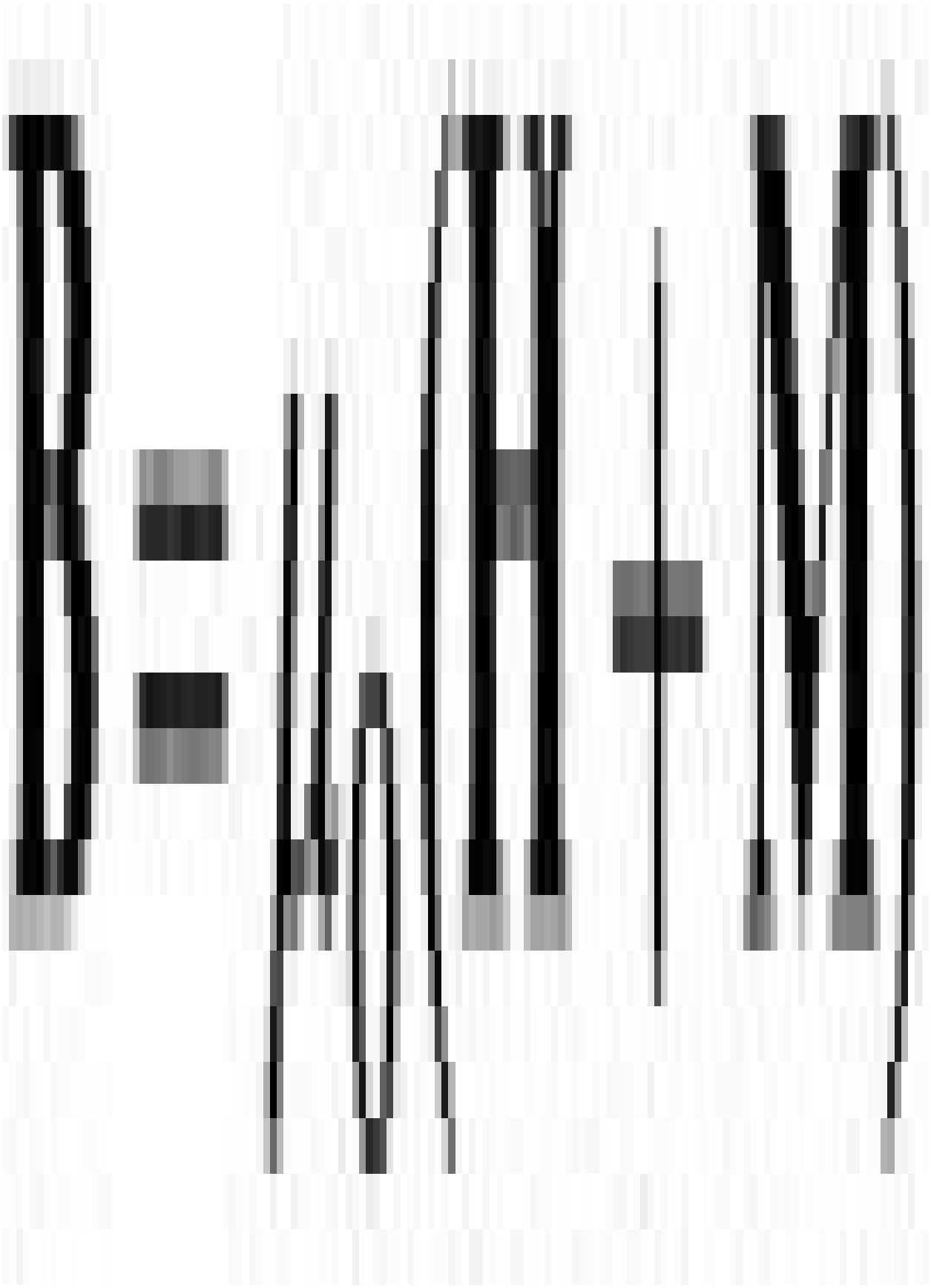
100 A. Superconducting magnets can now be made to produce fields of up to about 10 T within solenoids of lengths up to about 1 m and inside diameters up to about 40 cm. A 10 T field would be given by 100 A flowing in 8×10^4 turns per metre of wire. Higher fields can be obtained using filaments of niobium–tin alloy which can carry larger currents and remain superconducting. New materials which are superconducting at higher temperatures are continually being fabricated. Some oxides containing copper, barium and a rare-earth element have been found to be superconducting at temperatures as high as 125 K, allowing them to be used with liquid nitrogen cooling, which greatly reduces cooling costs. If the resistivity of these materials stays essentially zero for long enough time at large current densities and high fields, and if wires or other useful conductors can be made out of them, they will become commonplace in magnetic material applications.

Superconducting magnets without iron offer the added advantage that the relationship between field and current is linear, since there are no hysteresis effects. Precautions must be taken in their use to avoid the effects of a sudden loss of superconductivity from whatever cause. The rapid deposition of 100 A into large lengths of rather thin wires of normal conductor would cause severe heating problems.

5.4.2 Permanent magnets

As discussed above, if a piece of steel is magnetized in an external field and then removed from the field it remains magnetized. Its magnetism can only be removed by an external demagnetizing field or by heat treatment. We will not go into the microscopic reasons why the domains in the permanent magnet arrange themselves so as to keep the magnetization. The description of ferromagnetism and the different behaviour of different ferromagnetic materials is left to courses on the solid state of matter. The occurrence of permanent magnetism is of great practical importance as it enables us to provide magnetic fields without power consumption.

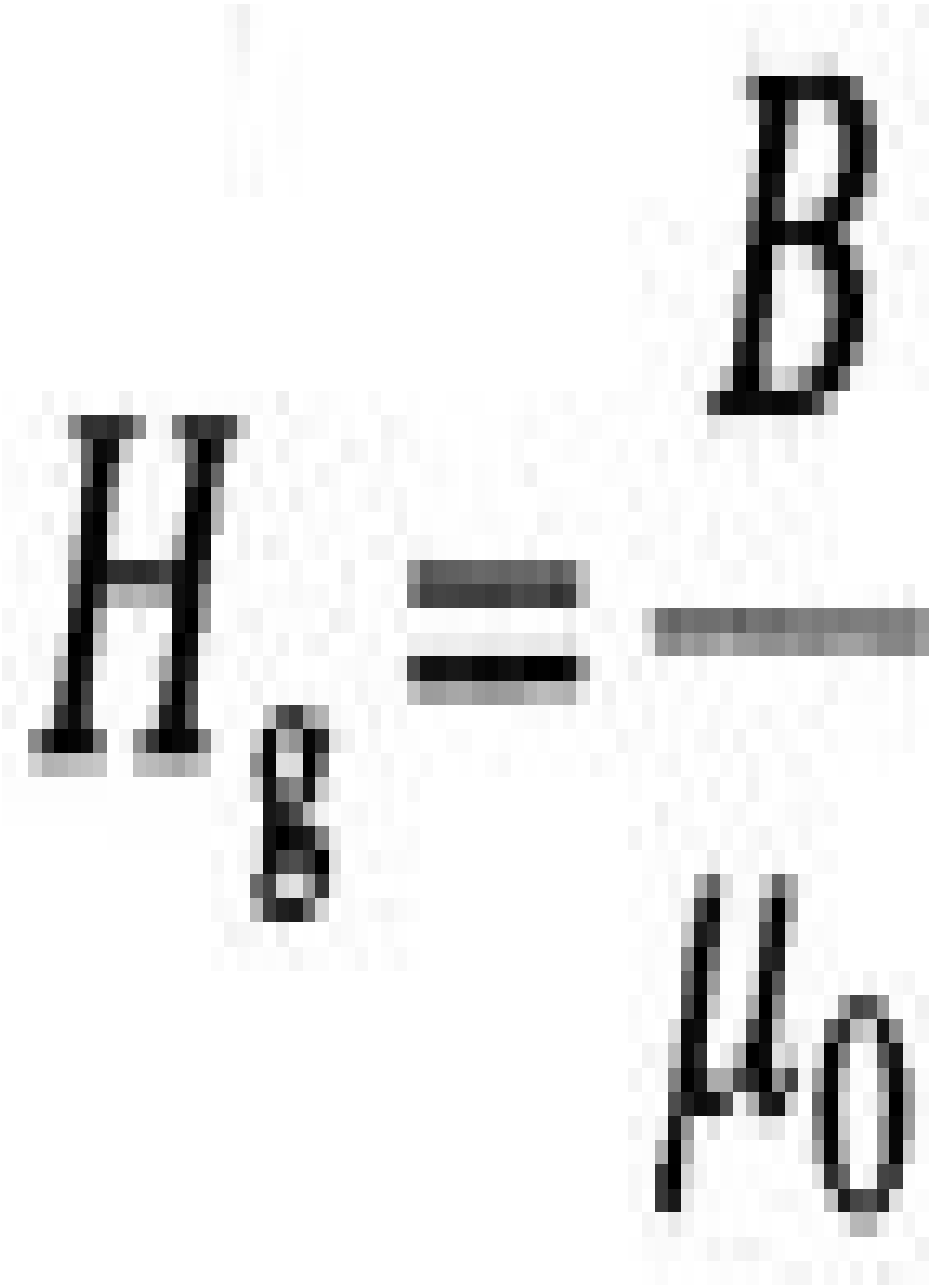
The chief point to grasp when considering fields due to permanent magnets is that the magnetization M of the material does not arise by virtue of an external field. If there are no external fields the material is still magnetized, and the magnetization itself acts as the source of the fields B and H. These fields obey the usual boundary conditions, and the relation



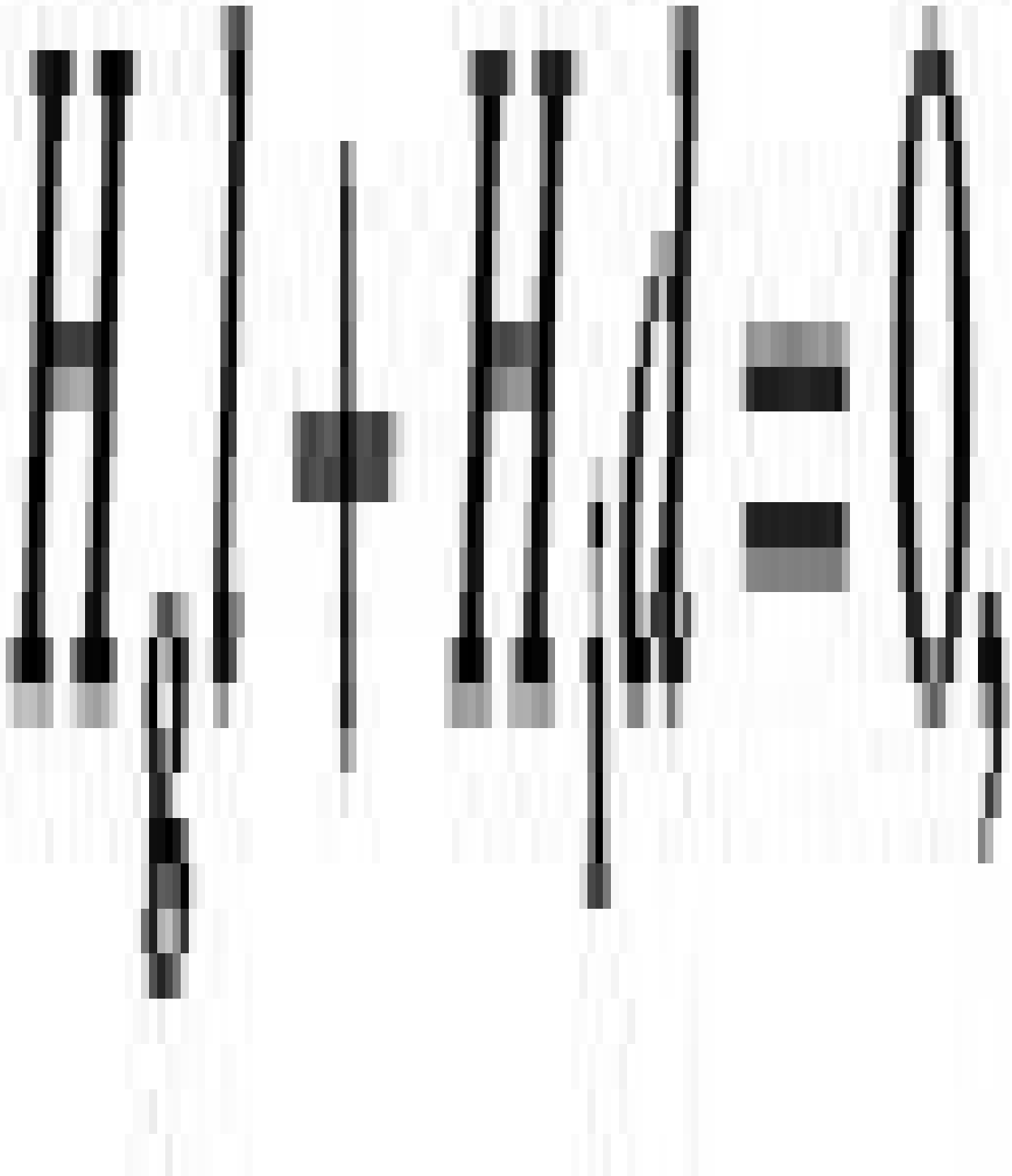
still holds at every point. The boundary conditions that the magnetic fields obey impose certain limitations on the distribution of magnetism within a permanent magnet. For example, a piece of material which has straight cut edges with sharp corners cannot be uniformly magnetized. The boundary conditions would imply a discontinuity in the field outside the sharp corners, and this is not allowed. However, the edge effects are rather small, and permanent magnets are uniformly magnetized throughout almost all of their volume.

What is the field around a permanent magnet consisting of a rod of (nearly) uniformly magnetized material? There are no free currents flowing, and if we take the line integral $H \cdot dl$ around a closed loop passing through the magnet, then $H \cdot dl = 0$. This implies that the magnetic intensity inside the rod must be in a direction opposite to that outside, and opposite to the direction of the field B, since the lines of the field B are continuous. The lines of the fields B and H of the magnetized rod are shown in Figure 5.20. Compare this drawing with Figure 5.10. At distances from the rod much greater than the length of the rod the field looks like the field of a dipole. The field far away from a magnetized specimen of any shape looks like a dipole field, since the more complex parts of the total field fall off more rapidly with distance than does the dipole part. The magnitude of the dipole field is determined by the magnetic dipole moment of the rod, which is simply its volume times the magnetization M.

The magnitude of the field generated by any permanent magnet depends on the magnetization of the magnetic material, and the magnetization is affected by the shape of the hysteresis curve of the magnetic material and the geometry of the magnet. Consider a C-shaped permanent magnet. As we did when working out the field of an electromagnet we shall assume that the lines of the field B pass straight across the gap between the pole faces. The field B is normal to the pole faces and is continuous across their surfaces. Its magnitude B in the gap is thus the same as its magnitude in the yoke of the magnet. The magnetic intensity Hg in the gap is thus given by

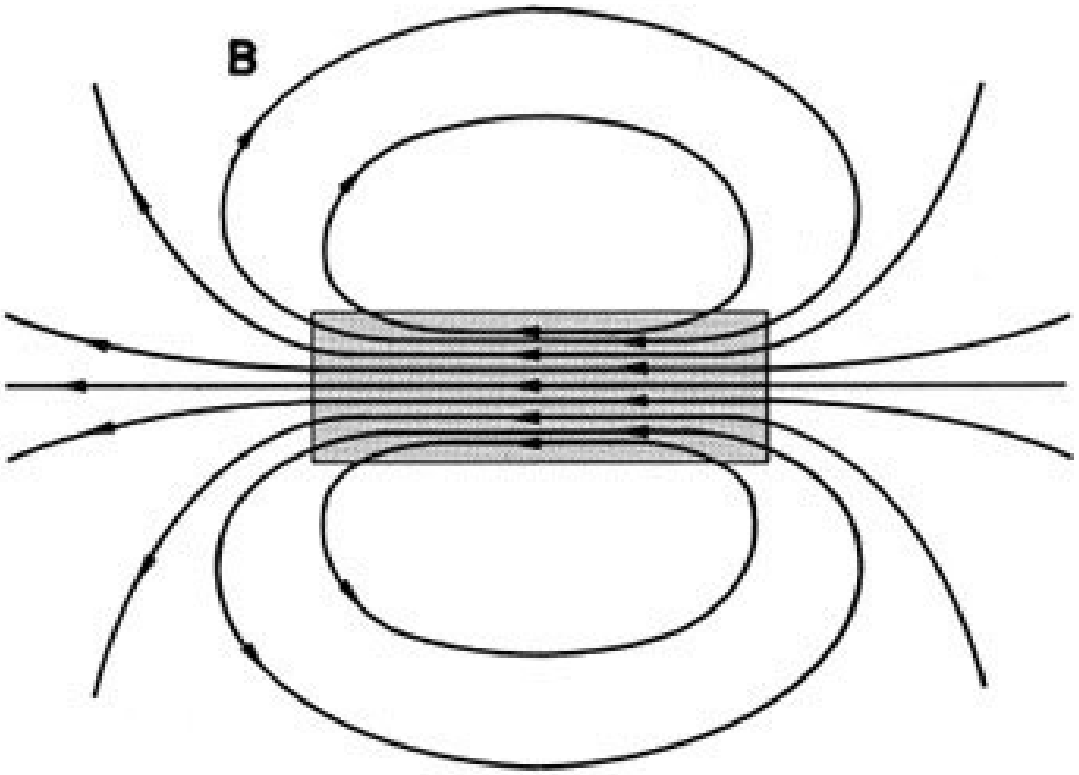


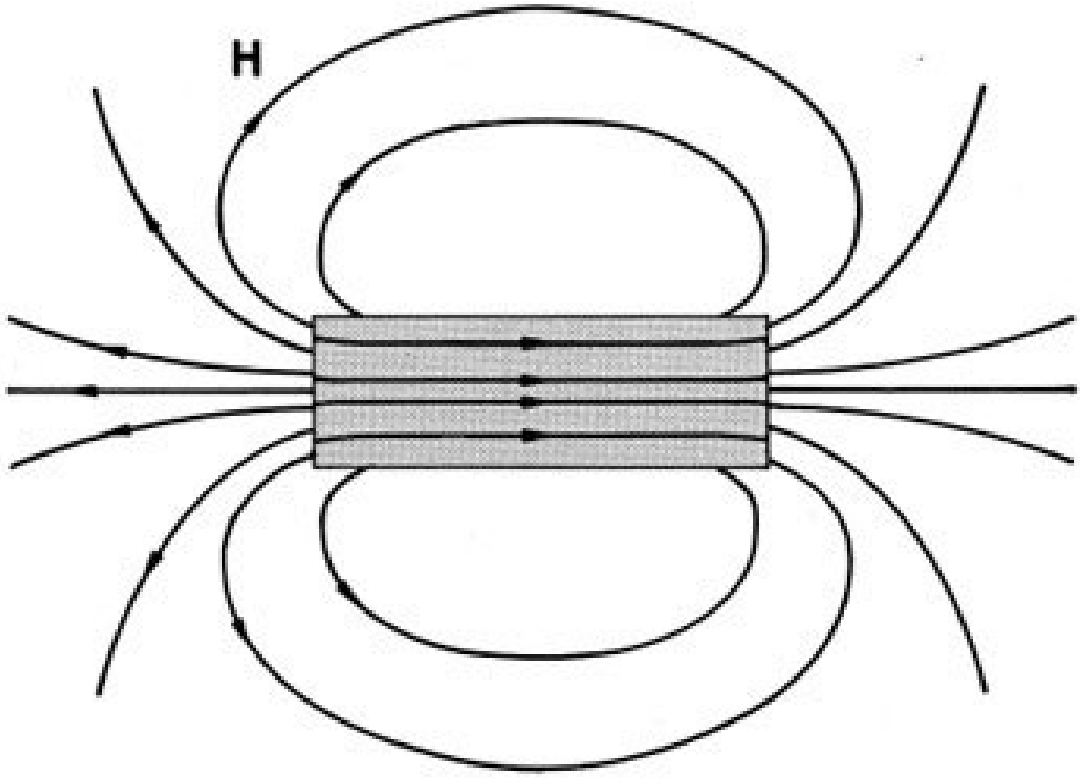
Since there are no free currents flowing the line integral of the field H over any complete path is zero, and



where l and d are the lengths of the gap and yoke respectively. Hence the field

Figure 5.20. The magnetic fields of a nearly uniformly magnetized bar magnet.


B


H

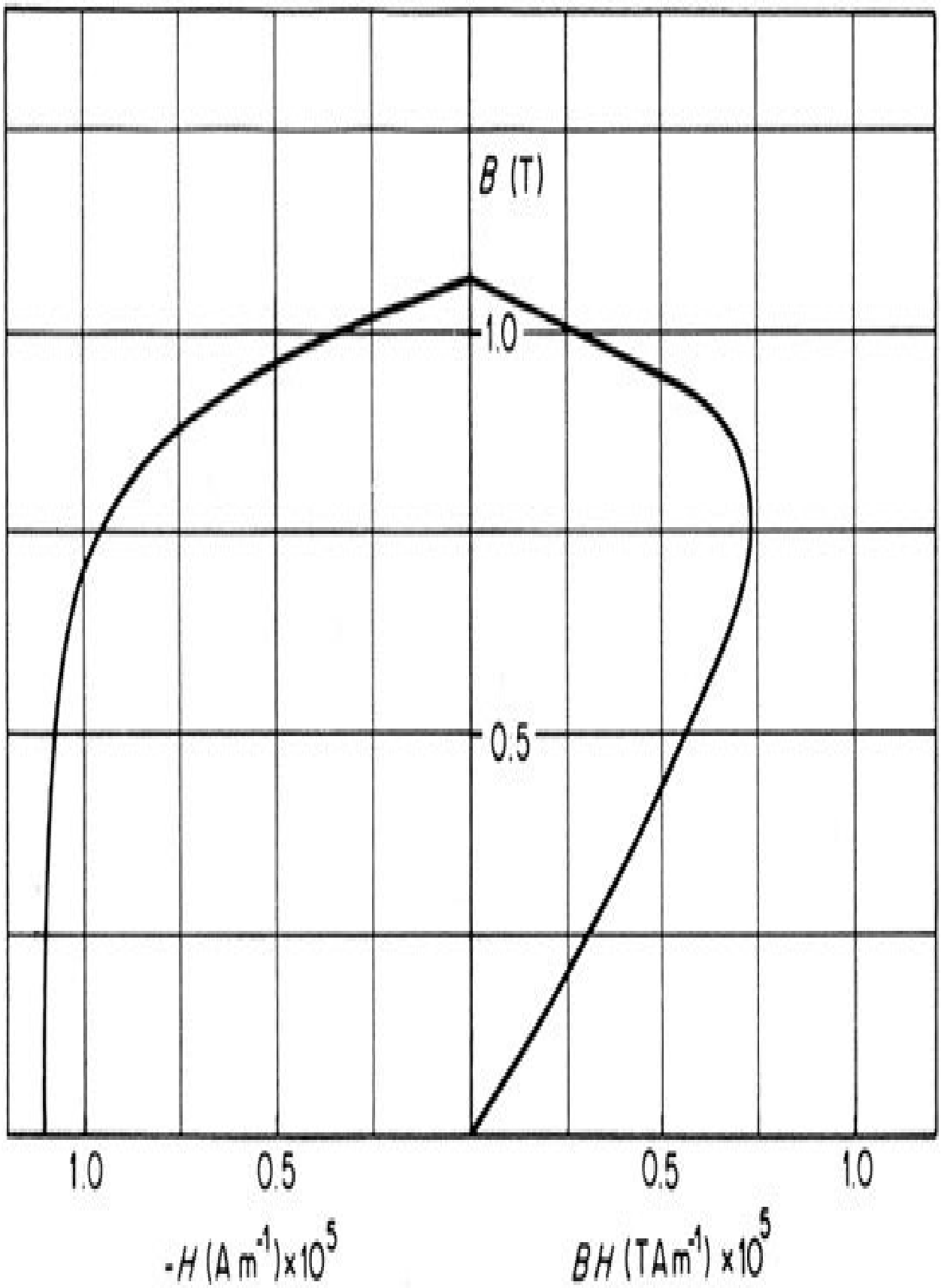
Hi in the yoke is

(5.46)

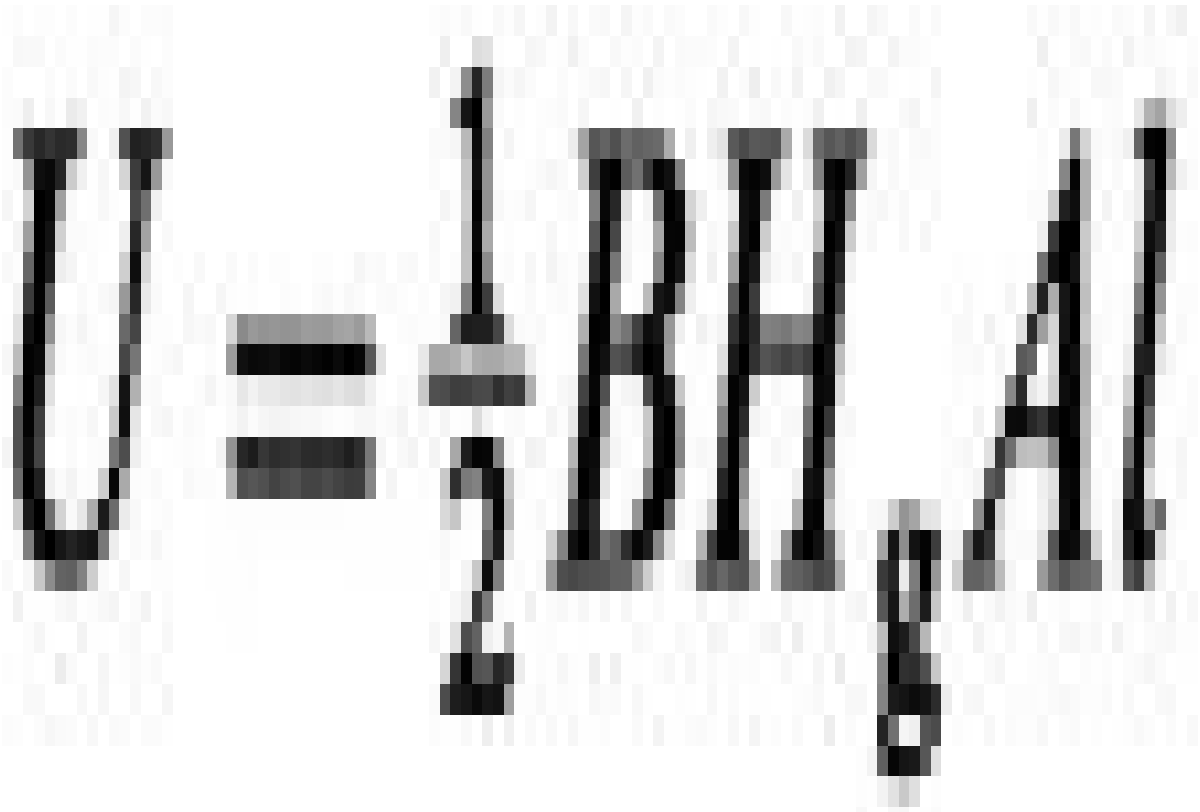
The values of B and H in the yoke must lie on the line defined by this equation.

The highest magnetic field B in the material of the magnet occurs when the material is in a state corresponding to some point on its major hysteresis curve. Figure 5.21 shows the relevant part of this curve for a commercial permanent magnet material. In our example the actual fields in the magnet are obtained by finding the intersection of the line given by Equation (5.46) with the B–H curve. For a gap length of 2 cm and length of yoke 30 cm, the field in the gap obtained in this way is equal to about 1T. The actual field in the gap will be somewhat less than this due to flux leakage.

Figure 5.21. The hysteresis curve for a typical commercially available permanent magnetic material. The appropriate quadrant of the B–H curve is shown together with the curve showing how the BH product varies with the field B.


B (T)
1.0
0.5
1.0
0.5
0.5
1.0
-H (A m-1) x 10^5
BH (TAm-1) x 10^5

As can be seen from Figure 5.21 the maximum field obtainable, whatever the lengths of gap and yoke, is about 1T. Normally one is interested not in obtaining the highest field but in obtaining a reasonably high field using the smallest volume of material. This is done when the ratio of the field energy stored in the gap to the volume of material used is a maximum. If the cross-sectional area of the yoke is A the energy stored in the gap is



from Equation (5.46). The volume of the yoke is Ad, hence using the material most efficiently corresponds to maximizing the BH product in the material. The maximum BH product of a substance is thus a measure of its ability to produce high fields with small volumes. The permanent magnet should be designed so that it operates at the point on the B–H curve corresponding to the maximum value of BH. On the right-hand side of Figure 5.21 the quantity BH is plotted against B for the material whose B–H curve is shown on the left-hand side. It can be seen that for optimum use of the material the field in the gap and yoke should be about 0.8 T, when the BH product has its maximum value of about 7×10^4 TAm^{-1}. Thus if the gap length of a magnet is to be cm, the length of the yoke should be about 14 cm.

5.5 SUMMARY OF MAGNETOSTATICS

A moving charge q experiences an electric force qE, independent of its state of motion, and also a magnetic force qv × B depending on its velocity v. The total force F is

(4.16)

and this equation defines the electric field E and the magnetic field B.

A steady stream of uniformly moving charges constitutes an electric current, and a steady electric current gives rise to a steady magnetic field. If a diagram of the field lines is made of any magnetic field B the lines are always continuous, and hence

(4.18)

over any closed surface S. This implies that at every point the relation

(4.19)

holds. The field B obeys Ampère’s law

(4.31)

where I is the current in a right-hand sense through the closed loop and μ_0 is the permeability of free space. This integral relation reduces to the differential relation

(4.47)

in terms of the current density j and the vector operator curl.

If matter is present in a magnetic field the atoms acquire an effective magnetic dipole moment $\langle m \rangle$. We can then define a macroscopic function M, called the magnetization of the medium,

(5.1)

where N is the number of atoms or molecules per unit volume. For isotropic paramagnetic ($\chi_B > 0$), or diamagnetic ($\chi_B < 0$), substances

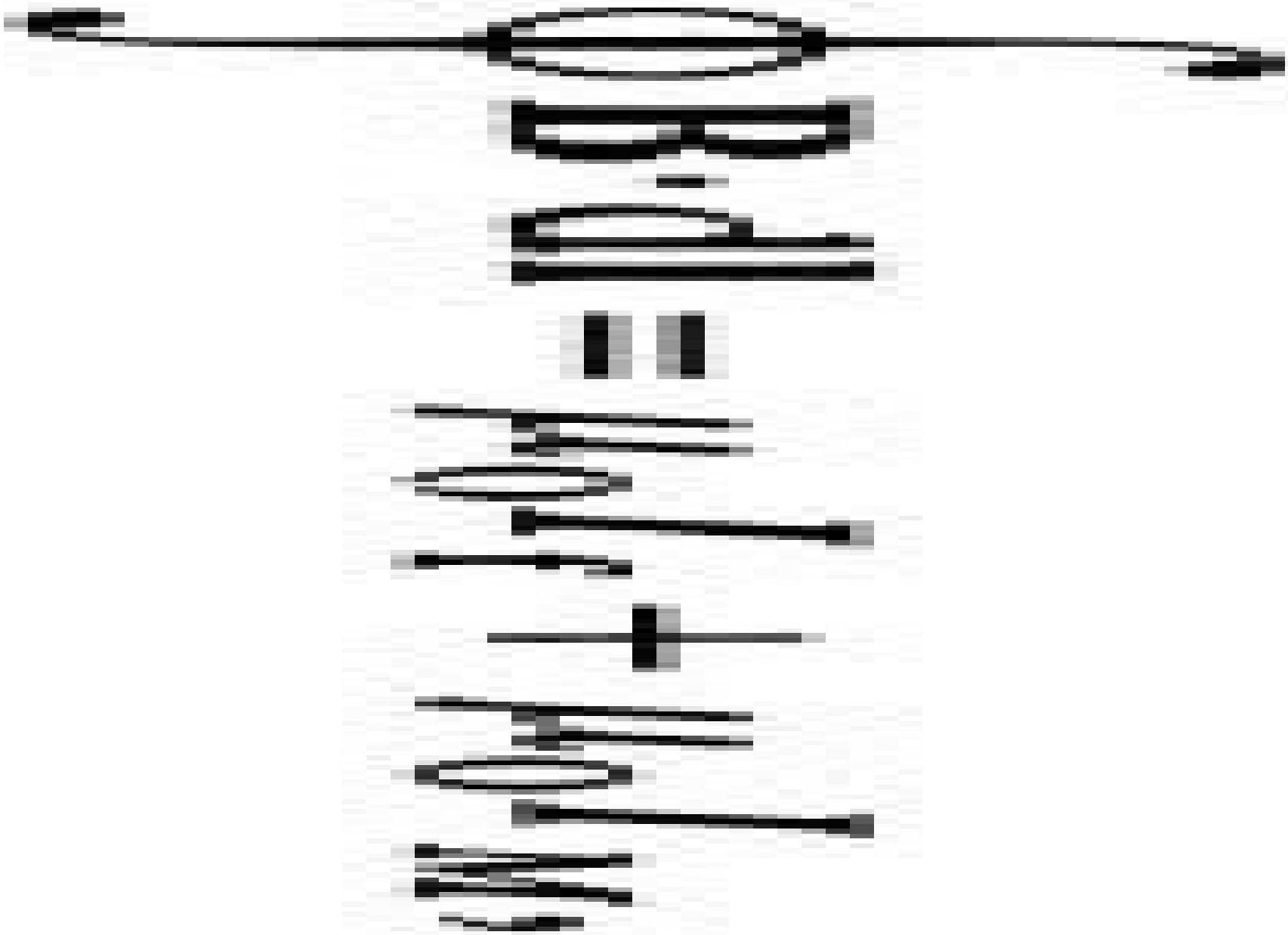
(5.23)

where χ_B the magnetic susceptibility, and B is the total magnetic field to which the magnetization of the material now contributes.

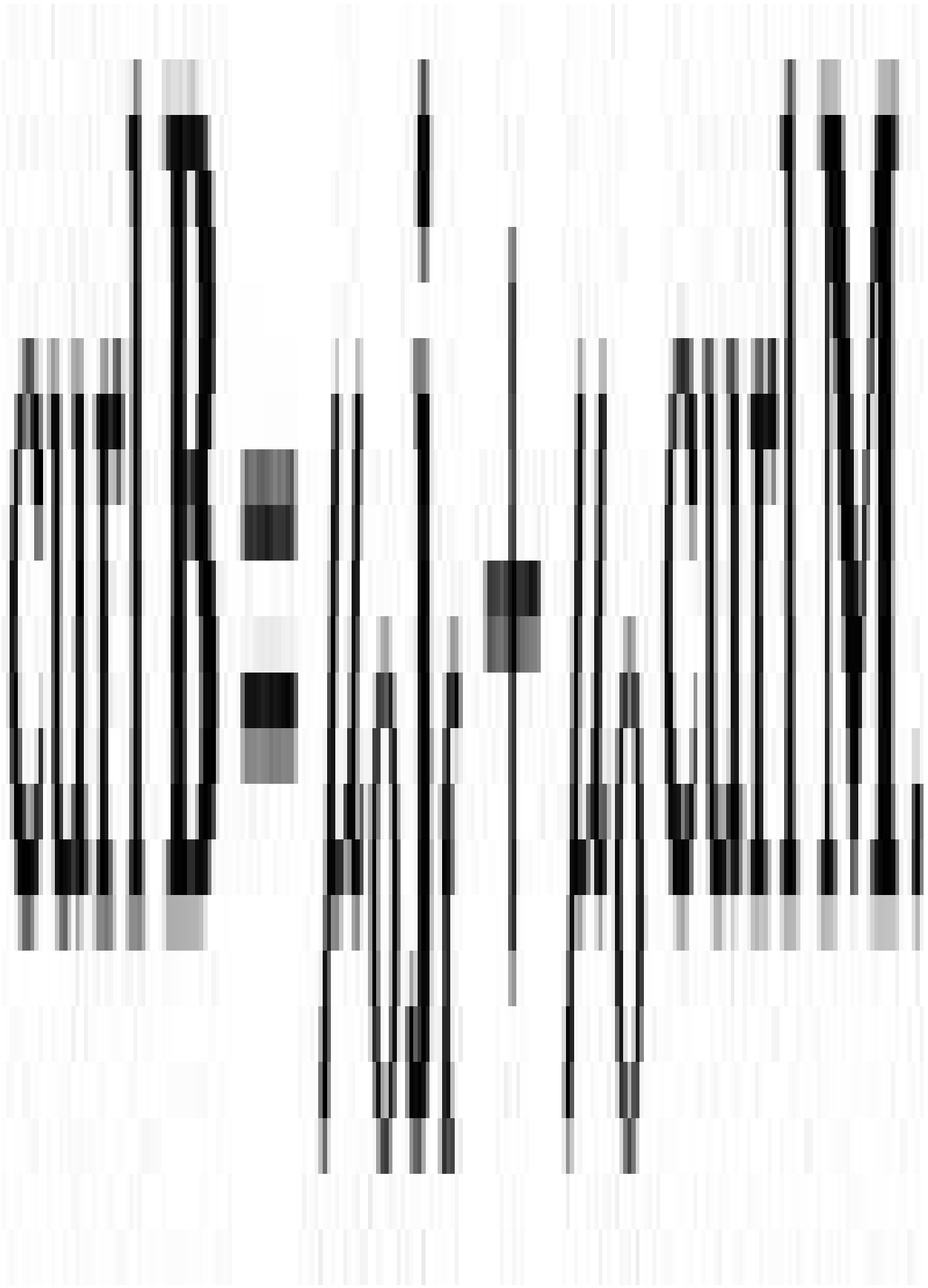
The magnetization of a sample gives rise to the same magnetic effects as a volume distribution of current within the sample given by

(5.19)

and surface currents over the sample. Over a closed loop Ampère’s law becomes, including the magnetization currents IM,



which reduces to the differential form,



This is an equation the fields B and M have to obey at all points.

Instead of working with the vectors B and M, the magnetic intensity H is introduced as an auxiliary vector.

(5.27)

whence

(5.30)

The magnetic field H obeys this equation everywhere. The fields B and H obey the boundary conditions

(5.37)

(5.39)

at boundaries between media. For non-ferromagnetic materials

(5.32)

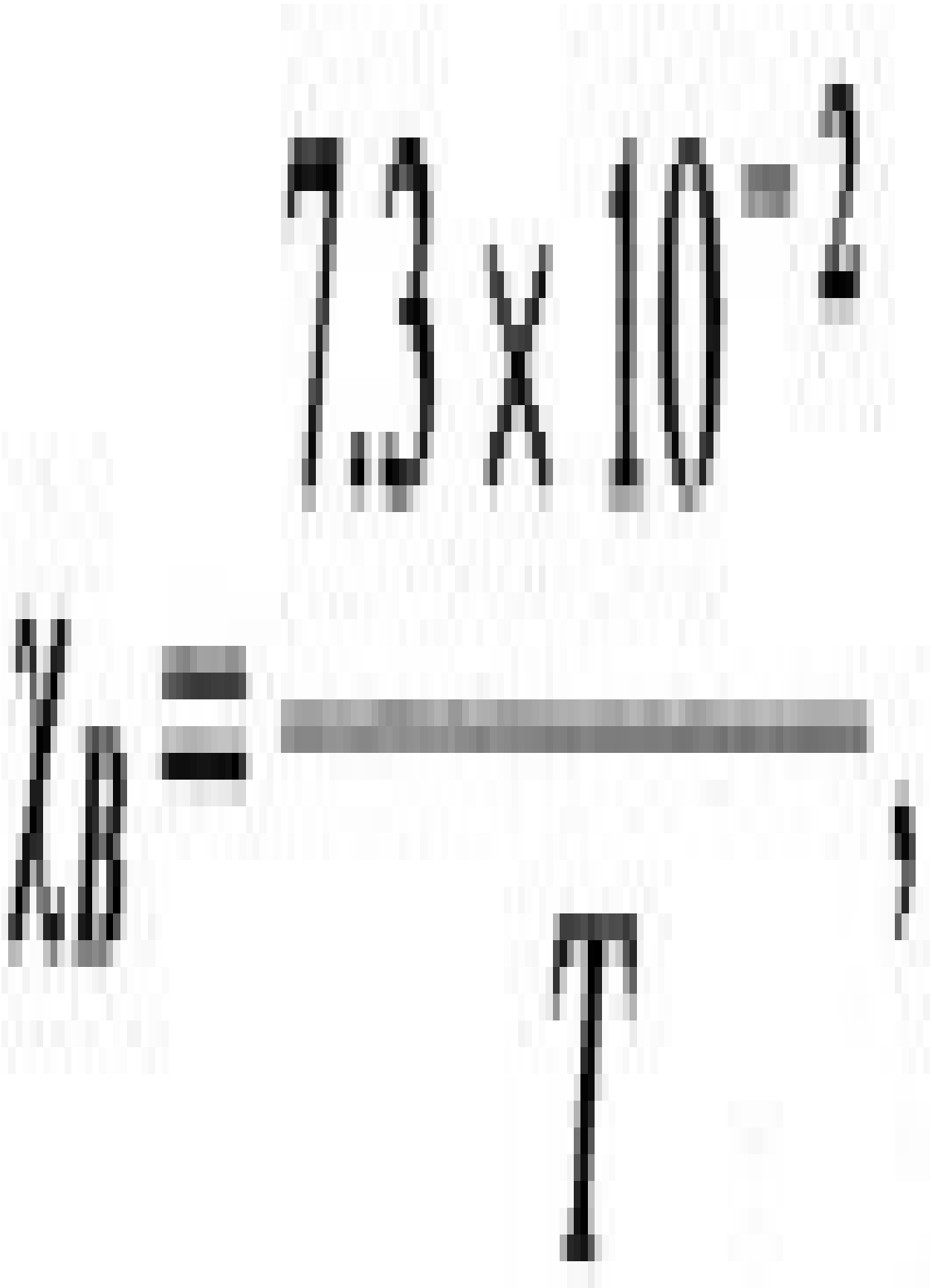
where the constant μ is given by

(5.33)

and is called the relative permeability of the material. For ferromagnetic materials, the relationship between B and H is non-linear, and for hard ferromagnetic it is multiple-valued as shown in hysteresis curves. Hard ferromagnetics may retain some magnetization in the absence of external fields and are used to make permanent magnets.

PROBLEMS 5

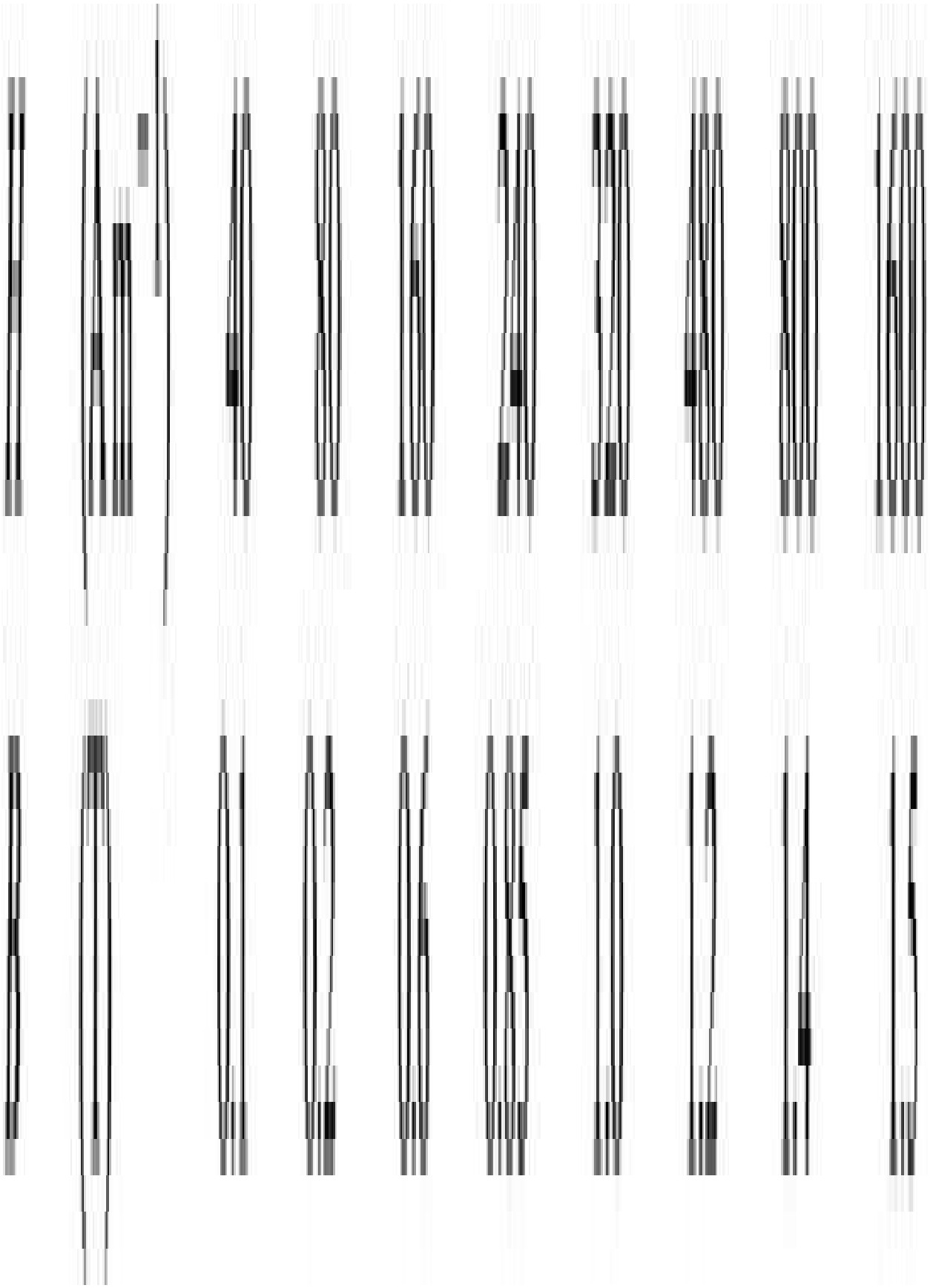
5.1 The magnetic susceptibility of a compound whose molecular weight is 400 and whose density is 2×10^3 kgm−3 is given by the formula



where T is the absolute temperature. Calculate the permanent dipole moment associated with each molecule. Make a rough estimate of the diamagnetic correction to the above equation at room temperature.

5.2 Nickel is ferromagnetic at ordinary temperatures. At very large magnetic fields it becomes saturated, and corresponding values of B and H are 1.275 T and 5×10^5 Am^{-1}. Calculate the magnetic moment per Ni atom in Bohr magnetons. The density of Ni is 9×10^3 kg m^{-3}.

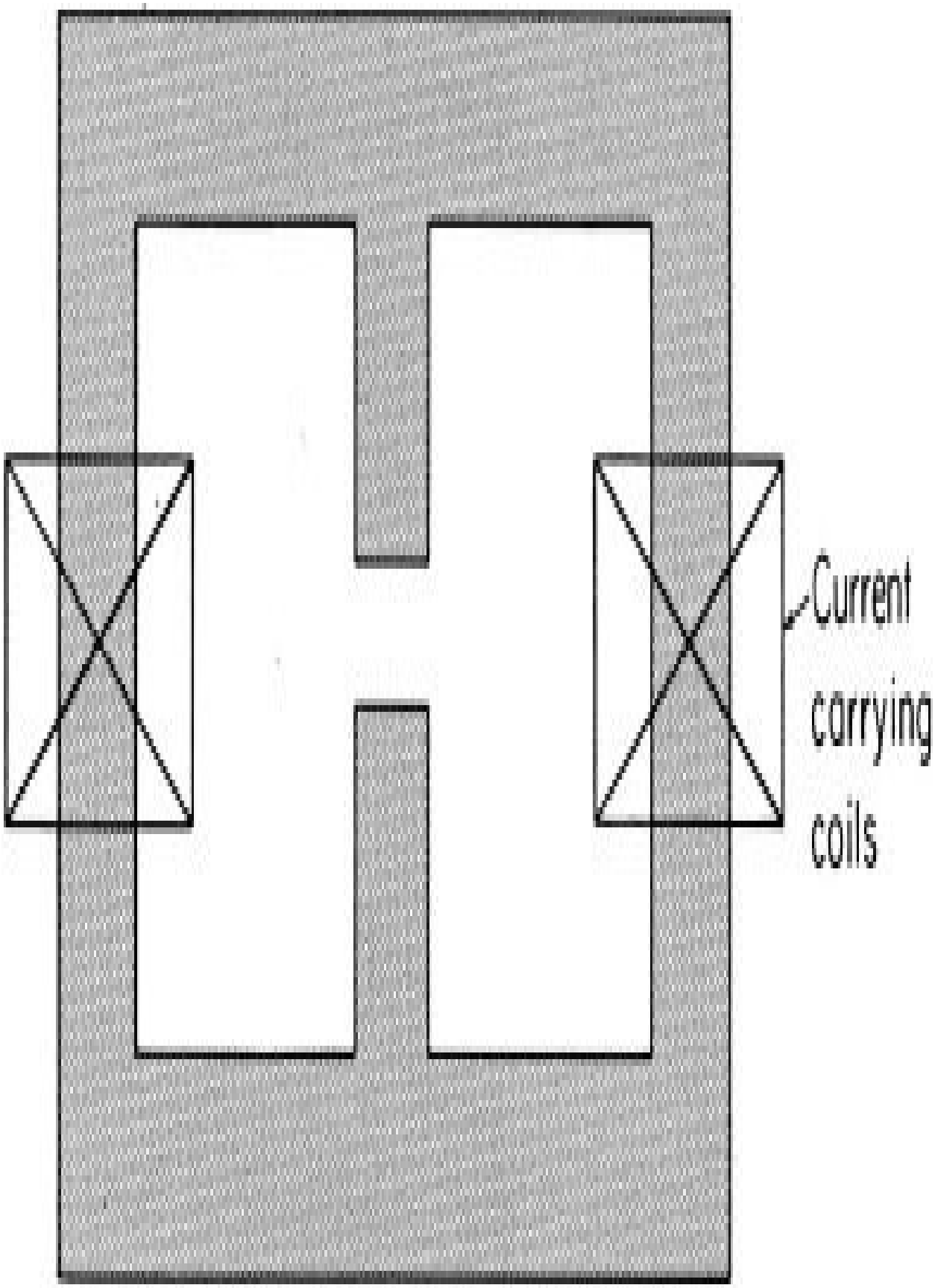
5.3 A small transverse gap is cut in a soft-iron anchor-ring (toroid) overwound with a coil carrying a current. The ring has a mean radius of 10 cm, the gap is 0.1 cm long and there are 200 turns of wire carrying 5 A. If the B–H relation for the iron is given by the table below, what is the field in the gap?



Compare the power required to maintain a field of 0.6 T in the gap with that required to maintain a field of 1.2 T.

5.4 A magnet core has the shape shown in Figure 5.22. The cross-section of the iron has the same shape and size everywhere. Two identical coils carrying the same current (in directions such that there is a large field in the gap) are wound on the arms of the yoke. If the current through one coil is stopped, but maintained at its original value through the other one, and the permeability of the iron is assumed to be constant, how does the field in the gap change?

Figure 5.22.


Current
carrying
coils

5.5 Sketch the lines of the fields B and H in the following situations:

(i) a sphere of material of large constant permeability placed in an originally uniform magnetic field,

(ii) an isolated permanent magnet in the form of a uniformly magnetized sphere.

Use an appropriate change of scale for the density of the lines of the two different fields.

5.6 A cylindrical specimen of iron of length 10 cm and diameter 5 cm is very nearly uniformly magnetized parallel to its axis. It has a dipole moment of 75 Am^2. Estimate the magnetic field B inside the iron.

5.7 The magnetic field B inside a superconducting material is ‘frozen’ at the value it had when the material became superconducting. A bar magnet is brought near to the horizontal surface of a rigid superconductor in which the magnetic field B is zero. Sketch the lines of the field B around the magnet and show that outside the superconductor they are the same as those due to the real magnet together with a fictitious image magnet located inside the superconductor. If the bar magnet is released make a rough prediction of its subsequent behaviour.

5.8 Estimate the magnetic field B at a point P on the axis of a long solenoid of length L, filled with a rod whose relative permeability may be taken to have the constant value μ. The total number of turns on the solenoid is Nt and the current I. The point P is outside the solenoid, and the diameter of the nearest end of the solenoid subtends an angle 2β at P; the diameter of the farthest end subtends an angle 2α at P.

5.9 It is sometimes necessary to know the field of a uniformly magnetized permanent magnetic material of a given shape, for example a uniformly magnetized sphere. The method of determining the fields in these situations is very similar to the solution of problems in electrostatics. Determine the field of a uniformly magnetized sphere by following the treatment of a dielectric sphere in a uniform electric field discussed in section 3.3.2 and using the potential ψ introduced in Problem 4.14.

5.10 A permanent magnet is in the shape of a toroid of diameter 10 cm. It has a small gap cut in it of length 1 mm, and has magnetization 10^6 Am^{-1}. What is the

magnetic field B in the gap?

5.11 An electromagnet is made of a material which, over the range of field strengths of interest, has a relative permeability μ equal to k (1 + a/B), where k and a are constants. The yoke length is L and it has a small gap of length l. The yoke is overwound with N turns of current-carrying wire. If L = kl determine the ratio of the power dissipated in the coils when the field B in the gap has size 2a to that dissipated when the field is equal to 3a.

*There may be a permanent magnetic dipole moment arising from the contribution of the nucleus or nuclei of the atom or molecule. This will be very small, and we will ignore it.

*See e.g. Surprises in Theoretical Physics, by R. E. Peierls. Princeton University Press, 1979.

*This definition of magnetic susceptibility, relating the magnetization to the magnetic field B which produces it, is a logical and transparent way to proceed with the development of magnetism. However, a different definition is often used, one relating the magnetization to the field H introduced in section 5.3. The numerical difference between the two definitions is small except for ferromagnetic materials, and for these the magnetization curves are more useful in discussions of magnetic properties in any event. See the comment at the end of section 6.4.2.

* However the approximation of infinite current density is often made for normal conductors in order to simplify the mathematics whilst retaining the essential features of a problem.

CHAPTER 6

Electromagnetic induction and magnetic energy

The last two chapters dealt with magnetostatics—the magnetic effects which arise when steady currents flow. In this chapter we discuss what happens when currents are changing with time. Non-steady currents give rise to non-steady magnetic fields. It is an experimental observation that they also give rise to electric fields. This phenomenon is called electromagnetic induction, and the associated electric fields are called induced electric fields. In the following sections we will relate the electric field to the changing magnetic field and examine some of the consequences of the relationship.

6.1 ELECTROMAGNETIC INDUCTION

In section 4.1.1 we introduced the concept of electromotive force (e.m.f.). A battery has an e.m.f. of V volts if it maintains a constant potential difference V between its terminals. If the terminals are connected by a conductor of resistance R a current equal to V/R flows around the circuit completed by the battery and the external resistance. The electric field is everywhere static and the line integral of the field E over any closed loop is zero. The line integral of the field E over any path between two points A and B is constant, and is equal to the potential difference between the two points. If the two points A and B are the positive and negative terminals of the battery, the e.m.f. of the battery is

(6.1)

When a current flows in a wire joining the terminals, the battery must do work to keep the potential difference between the terminals constant. If a charge q moves from terminal A to terminal B around the external conductor, the work done by the battery is Vq:

$$V_q = q \int_A^B \mathbf{E} \cdot d\mathbf{l}$$

$$= \int_A^B \mathbf{F} \cdot d\mathbf{l}$$

where F is the force on the charge q. Hence

(6.2)

Equations (6.1) and (6.2) are equivalent definitions of the e.m.f. of a battery whose terminals are the points A and B.

We repeat that for an electrostatic field the line integral around a closed loop vanishes. If part of the loop consists of a resistive wire joining battery terminals A and B, the loop then being closed by going from B to A through the battery, the line integral is zero because the e.m.f. of the battery is balanced by the potential drop across the wire. When conditions are not constant with time there can exist an induced electric field for which the line integral around a closed loop is not zero. For this induced electric field, we may define the e.m.f. V around a closed loop or in a closed circuit in accordance with Equations (6.1) and (6.2) as

(6.3)

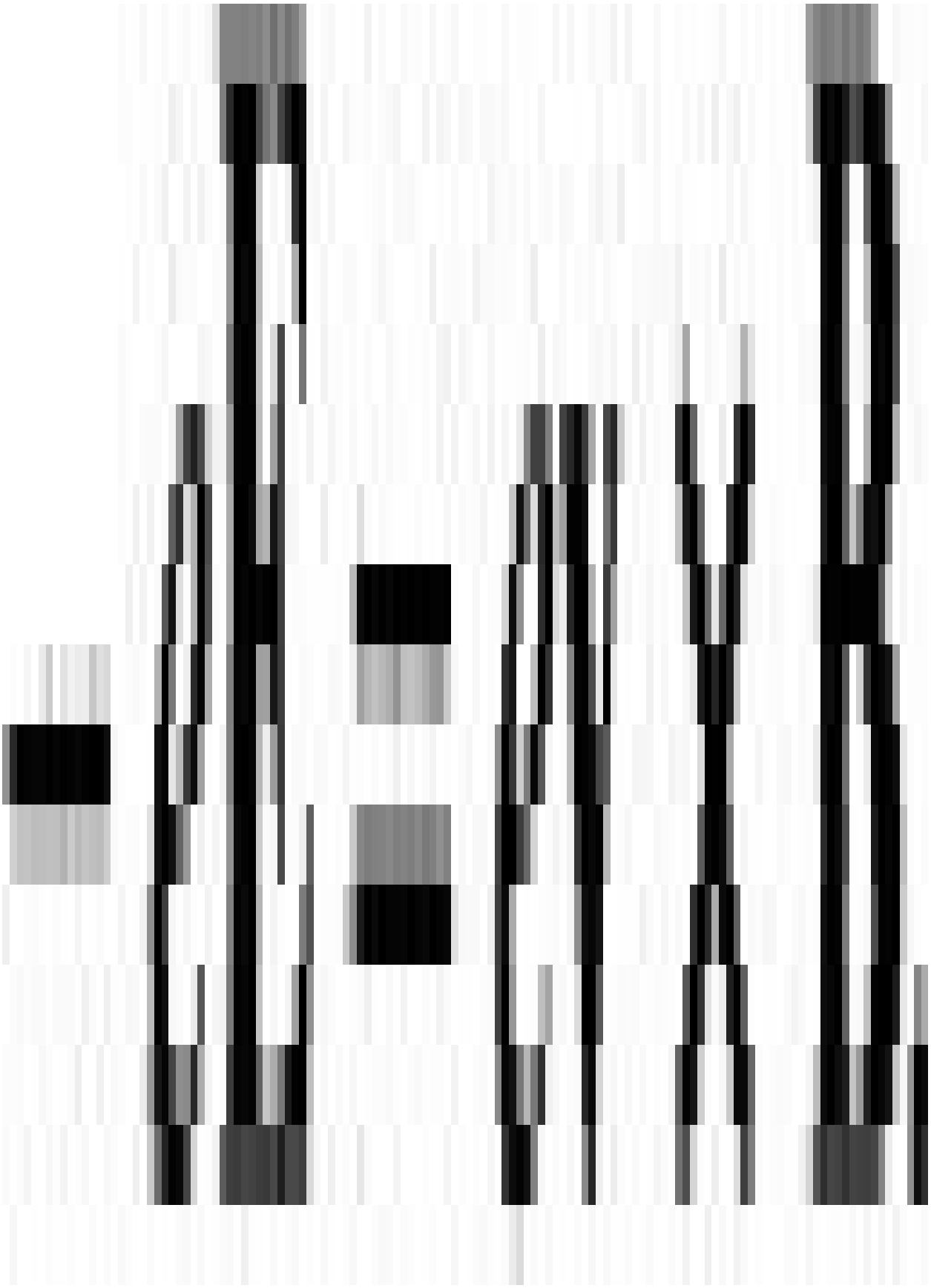
It should be noted that the field E in this equation is the total electric field and may include an electrostatic component, since this would not contribute to the line integral. If the closed loop is a conductor of resistance R a current I = V/R will flow around the conductor.

Two situations in which e.m.f.s exist in closed circuits are considered in this book. First we discuss motional e.m.f.s which are generated if part or all of a

circuit moves in the laboratory frame when a steady magnetic field is present. Motional e.m.f.s can be calculated using the Lorentz force law (4.15), and for a closed circuit can be related to the change of magnetic flux through the moving circuit. This relationship, obtained for motional e.m.f.s, remains true for the second case we consider, that of induced e.m.f.s generated when a circuit is stationary but the magnetic field in which it sits is varying with time, and we use the results of our considerations of motional e.m.f.s to introduce the law governing induced e.m.f.s.

6.1.1 Motional electromotive force

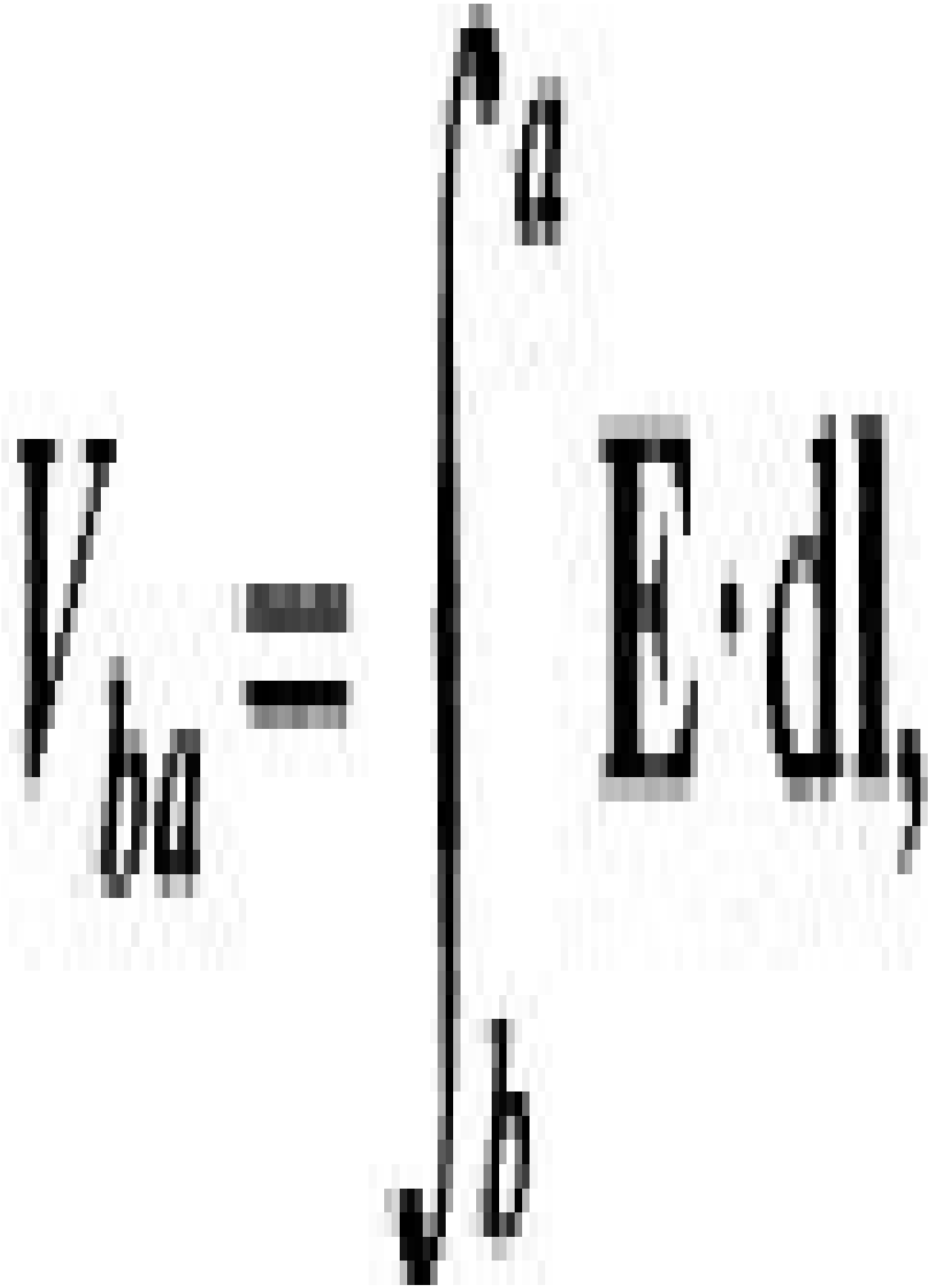
Consider first a metal rod moving with constant velocity v in a direction perpendicular to a uniform magnetic field B, as shown in Figure 6.1. There is a magnetic force given by Equation (4.15) on each electron and positive charge in the rod. The free electrons in the rod will move towards the end a and collect there, setting up a charge distribution which gives rise to an electric field which halts further migration of free electrons. The electric field set up, E, gives rise to a force – eE on the electrons exactly equal and opposite to the force – ev × B which acts on them because of their motion in the magnetic field, and hence



The electric field in the rod is thus given by

(6.4)

The potential difference Vba between the ends a and b of the rod is given by



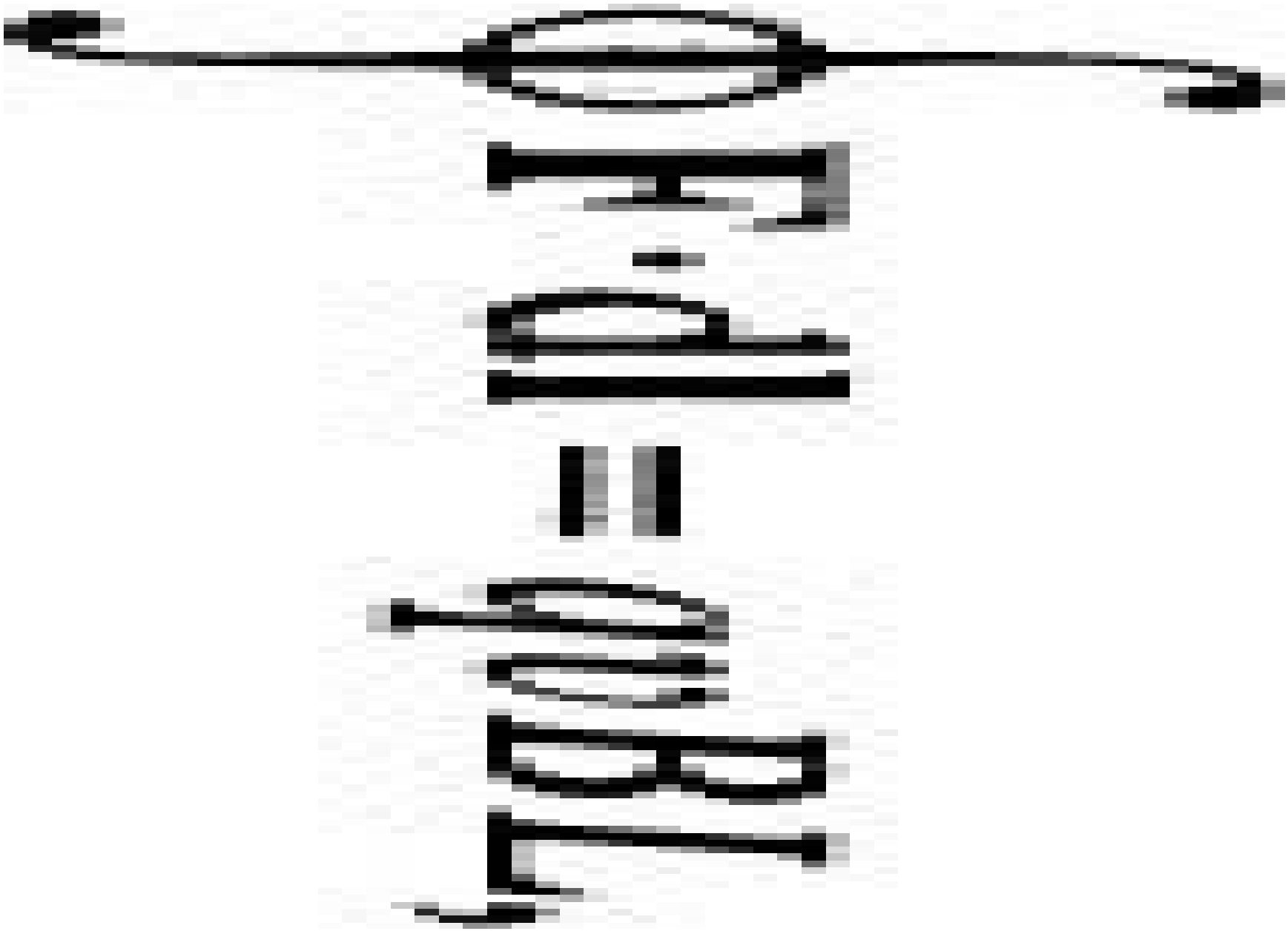
and hence

(6.5)

where L is the length of the rod.

As an example consider an aeroplane with a wing span of 50 m flying horizontally at 800 km per hour at a place where the vertical component of the earth's magnetic field is 2×10^{-5} T. There is a voltage difference between the wing tips, given by Equation (6.5), equal to 0.22 volts.

The potential difference between the ends of the rod shown in Figure 6.1 produces no current flow. If a metal rod moves on stationary conducting rails, perpendicularly to a uniform magnetic field, as shown in Figure 6.2, a current will flow. In this situation charges do not build up on the ends a and b of the rod; electrons traverse the complete circuit from b to a and around to b again, and continue to circulate. A stationary observer sees a magnetic force acting on the free electrons in the rod and this force pushes them over to end a. At the end a they push more electrons around the complete circuit, and a continuous current flows. The line integral of the force on a charge q evaluated around the complete circuit is given by



where L is the length of the rod. Since the net contribution to the line integral comes from the portion ab of the loop, the e.m.f. in the closed circuit is given by

(6.6)

Figure 6.1. A metal rod moving with velocity v perpendicular to a uniform field B.

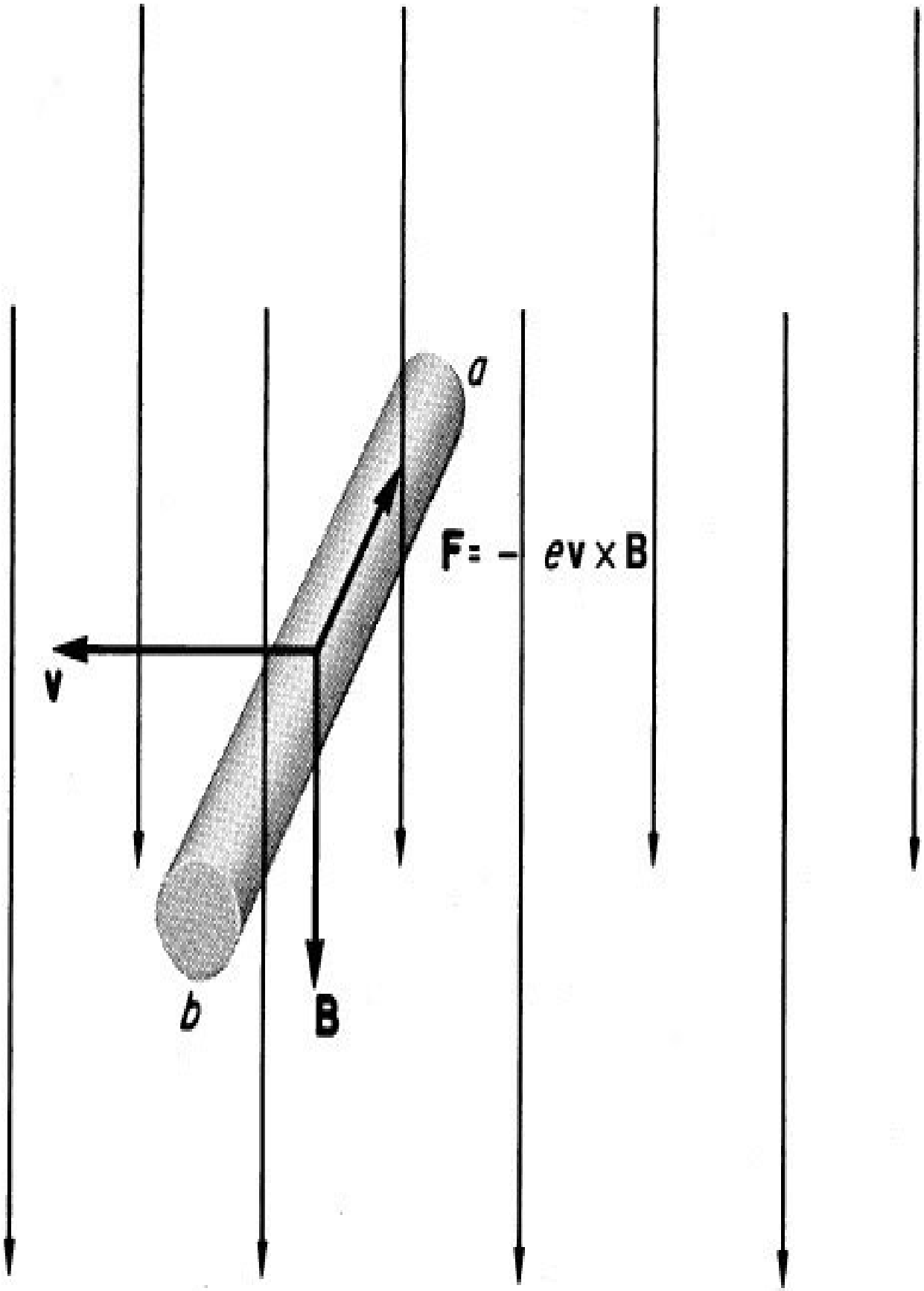
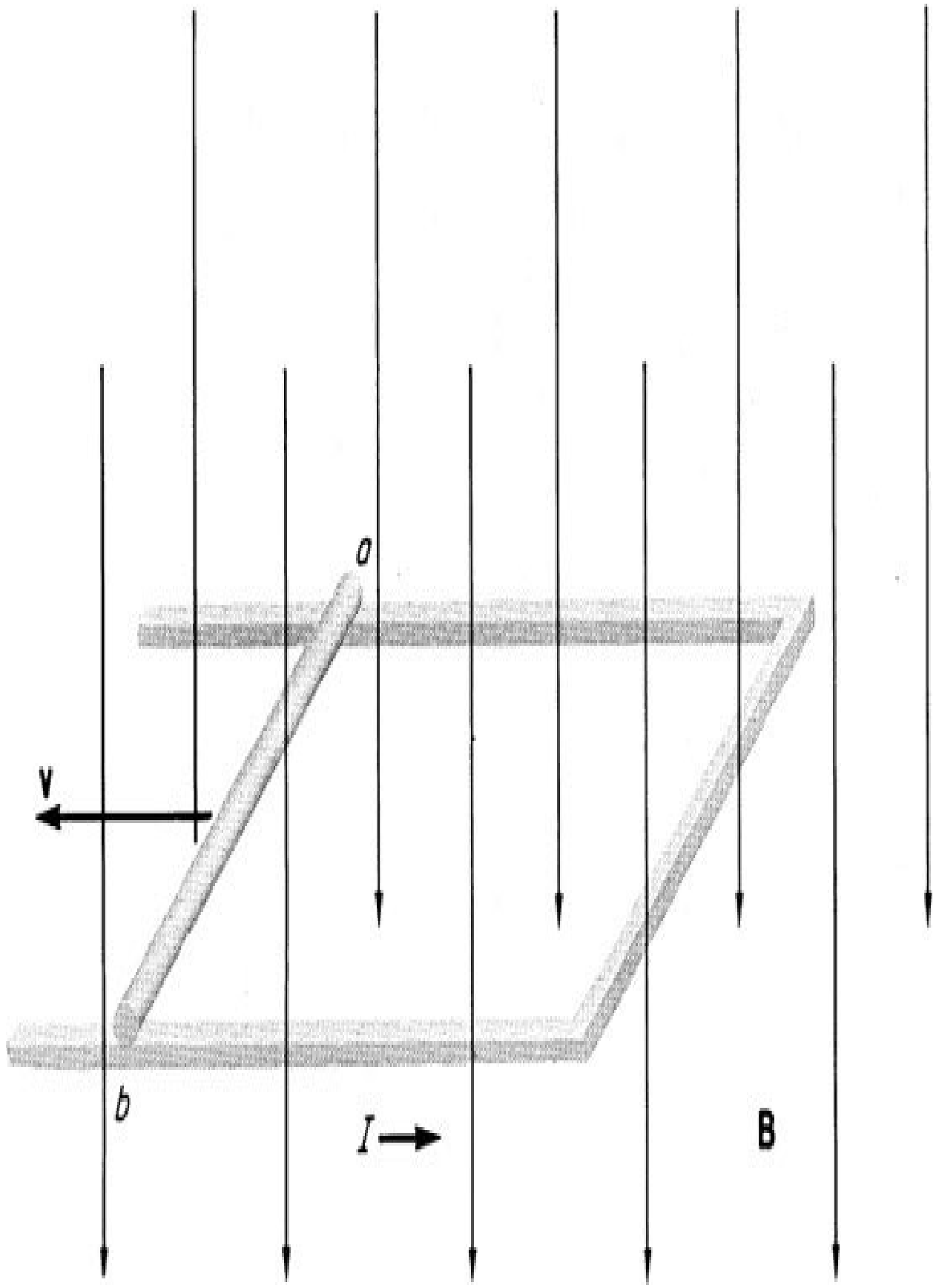

a
F = – ev x B
v
b
B

Figure 6.2. A metal rod moving with velocity v on stationary conducting rails. A uniform field B acts as shown.


a
v
b
I
B

This we may call a motional e.m.f. because it arises from the motion of a conductor in a magnetic field. The motional e.m.f. causes a current I to flow, given by

(6.7)

where R is the resistance of the circuit. The instantaneous power expended is $(vBL)^2/R$ and the work is derived from the mechanical force needed to move the rod along the rails.

Consider now the idealized situation shown in Figure 6.3, in which a uniform magnetic field B exists over a sharply defined region of space. A rectangular coil abcd moves in a direction perpendicular to the field B, with a speed v. While the coil is entering the field there is an e.m.f. of magnitude vBL in the circuit, and this e.m.f. gives rise to a current in the direction shown. When the whole coil is moving within the uniform field, the e.m.f. and current are zero, since the magnetic forces along the sides ab and cd cancel each other out. When the coil leaves the field there is again an e.m.f. equal to vBL but it now acts in the opposite direction. If we arbitrarily let the current be positive as the coil enters the field, the curve of current against time is as shown in Figure 6.4, where R is the resistance of the coil. The source of energy that keeps the currents flowing is the force that is moving the coil.

Figure 6.3. A rectangular circuit moving with constant speed v through a region of uniform magnetic field.

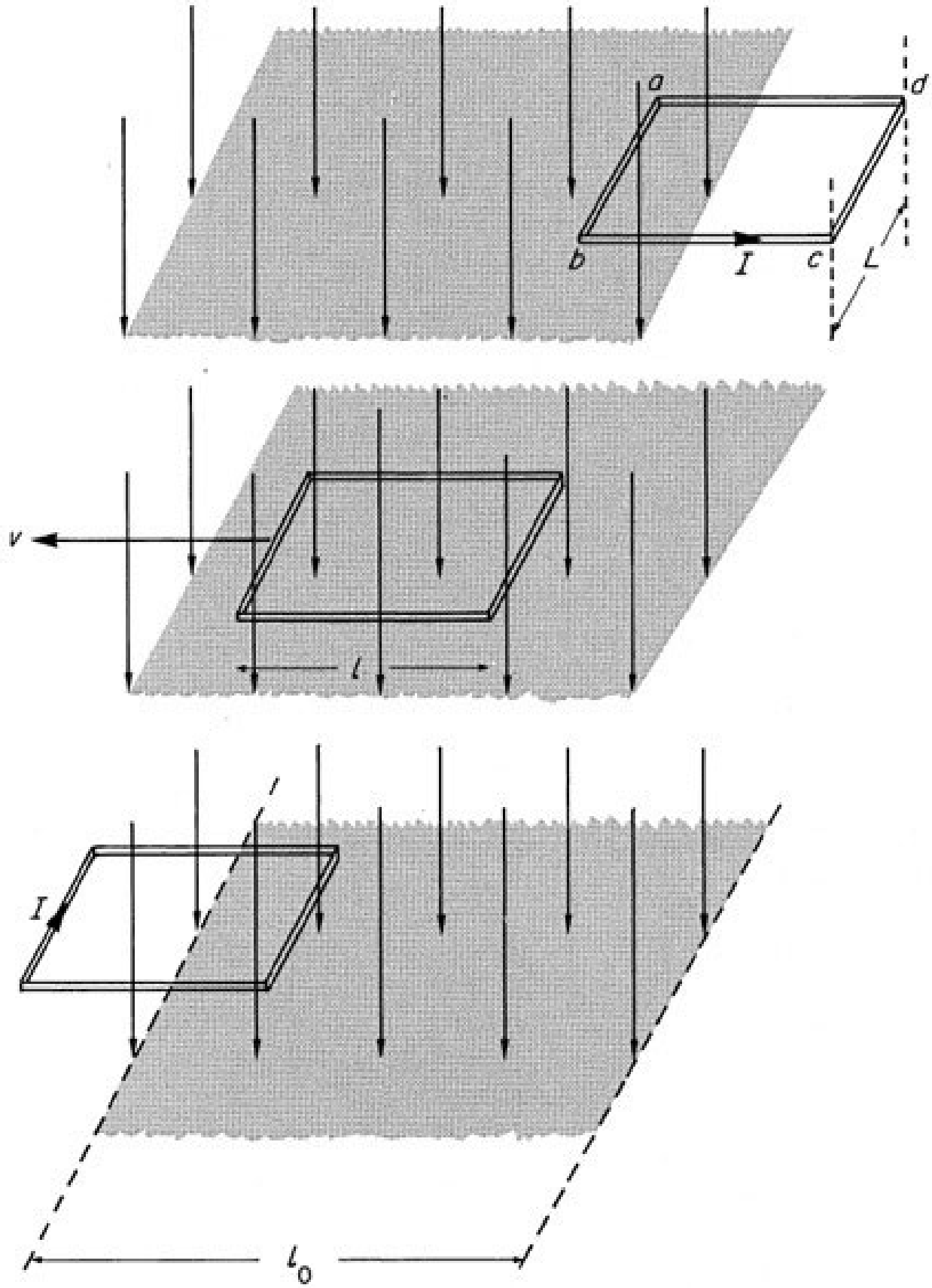
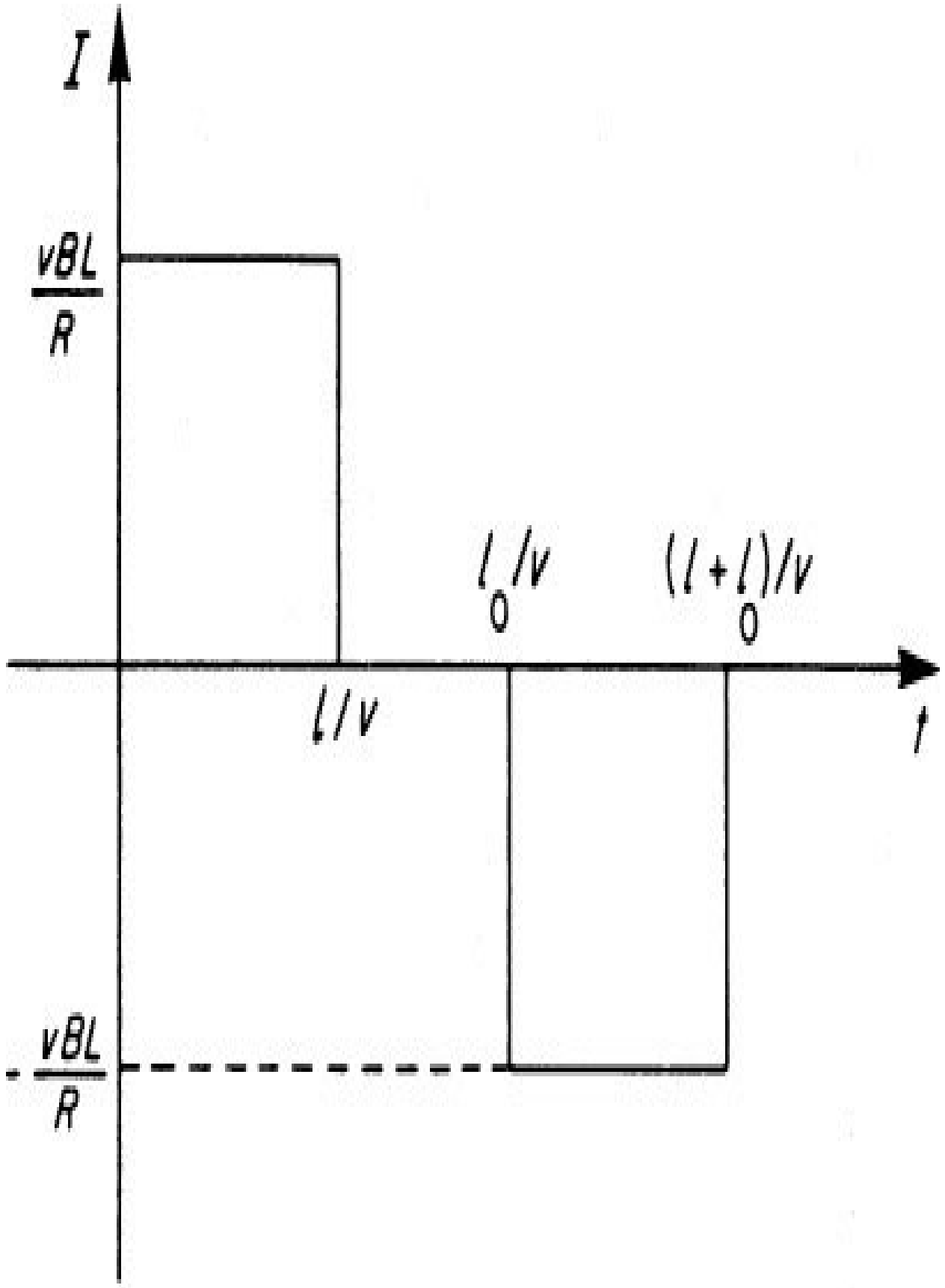

a
d
b
I
c
L
v
l
I
l_0

Figure 6.4. The variation of current with time for the moving circuit shown in Figure 6.3. The current is not exactly as shown; the induced currents do not change suddenly, but take some time to build up and die away.


I
vBL/R
l/v
l0/v
(l+l0)/v
t
-vBL/R

The e.m.f. in the coil as it enters or leaves the field can be expressed in terms of the total magnetic flux Φ through the coil. This flux is the sum of the part BLx due to the external field, where x is the length of the coil in the field, and the part due to the induced current in the coil itself. Since the induced current and the area of the coil are constant, this latter contribution to the magnetic flux is constant, and on entering the field

$$\frac{d\Phi}{dt} = BL\frac{dx}{dt}$$

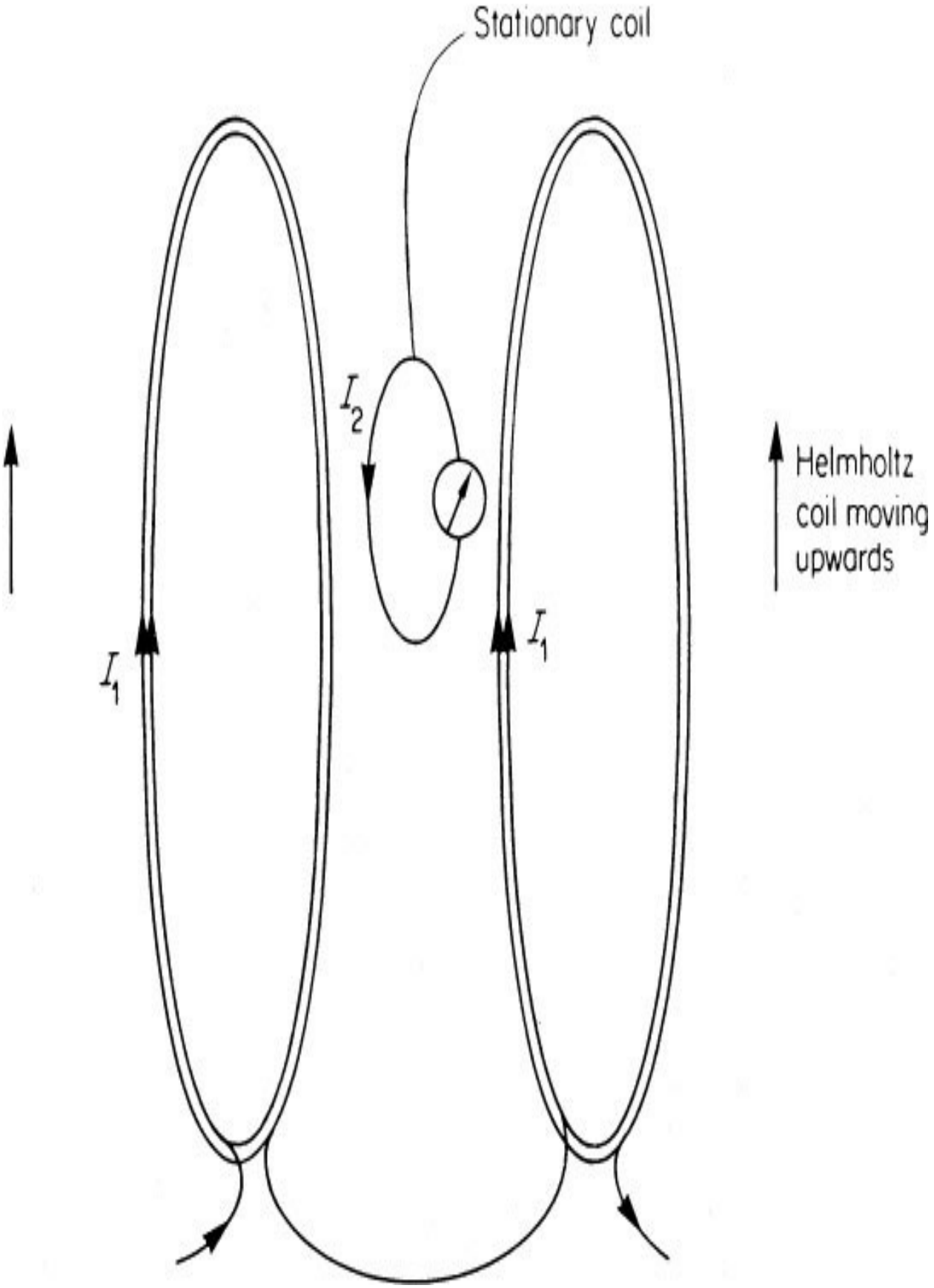
$$= BLv.$$

The magnitude of the e.m.f. is thus given by

(6.8)

From the point of view of an observer on the coil, the coil is stationary, and he observes a magnetic field sweep past him for a certain period of time. He measures a current flow in the coil during those times when the magnetic field does not cover the whole coil. Since the coil is stationary there is no net magnetic force on the free electrons, and he attributes the current flow to an electric field. Thus a static magnetic field for one observer looks like a magnetic field plus an electric field to an observer in uniform motion relative to the first*. Hence if a Helmholtz coil carrying a current I1 is moved past a loop of wire in series with a current meter, as shown in Figure 6.5, a current I2 will flow one way in the loop as the coil approaches, and then the other way as the coil recedes. Again the energy dissipated in the small coil as the currents I2 flow is supplied by the mechanical force needed to move the Helmholtz coil.

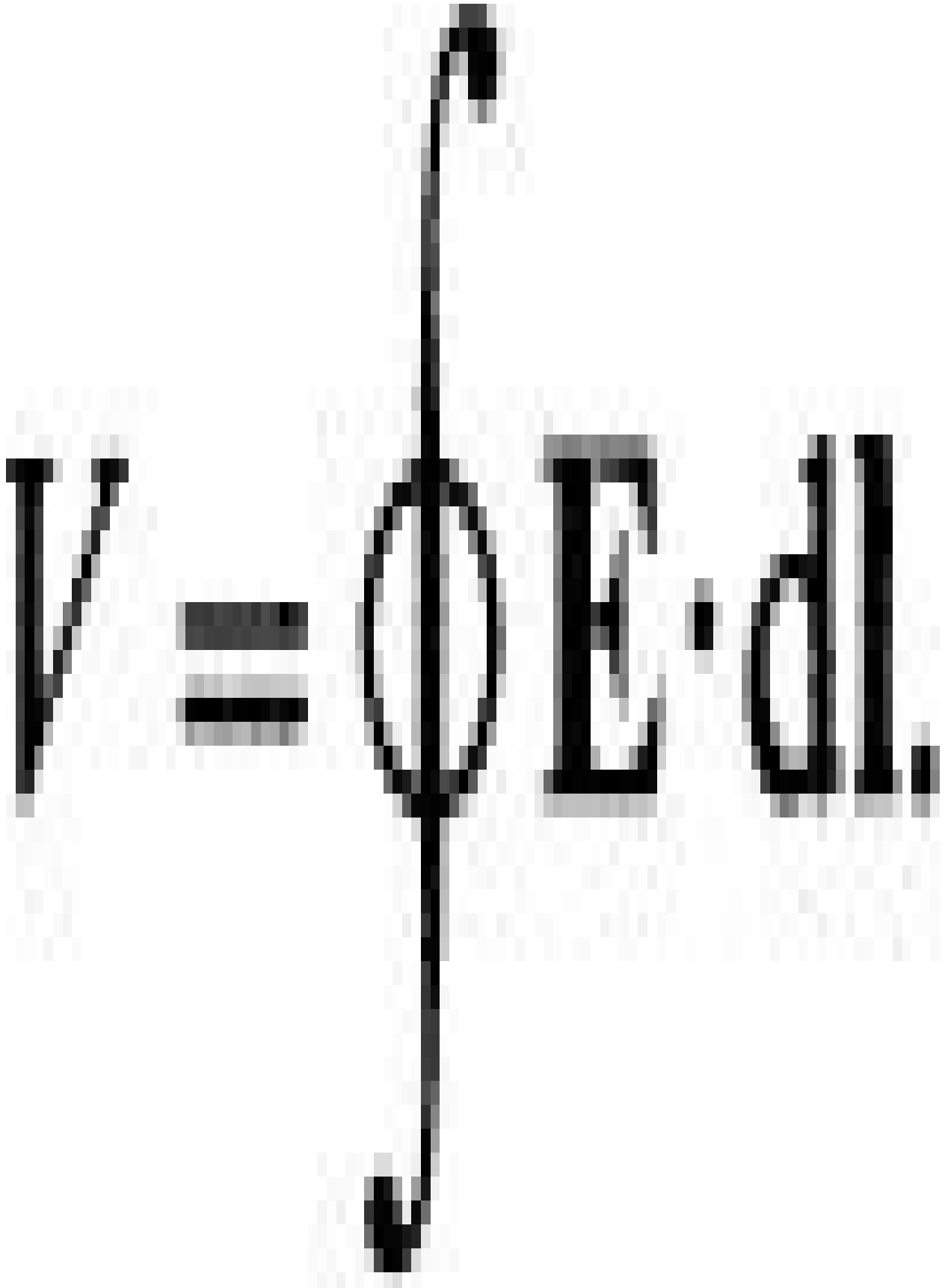
Figure 6.5. A Helmholtz coil carrying a current I1 moves past a stationary loop of wire. A current I2 is induced in the loop.


Stationary coil
I_2
I_1
I_1
Helmholtz coil moving upwards

6.1.2 Faraday’s law

We have seen that there is a motional e.m.f. in a circuit moving in a magnetic field in such a way that the magnetic flux through the circuit changes. An e.m.f. is also produced in a stationary circuit when the magnetic flux through the circuit changes because the sources of the magnetic field are moving. It is an experimental observation that an e.m.f. appears in a circuit when the magnetic flux through the circuit changes from any cause. Such an e.m.f. is called an induced e.m.f., and is related to the induced electric field caused by the changing magnetic field.

Consider two stationary coils, one carrying a current which gives rise to a magnetic flux through the second. If the current in the first coil changes there is an induced e.m.f. in the second coil which causes a current to flow. The induced e.m.f. V is equal to the line integral of the induced electric field E around the coil,


V = ∮ E · dl

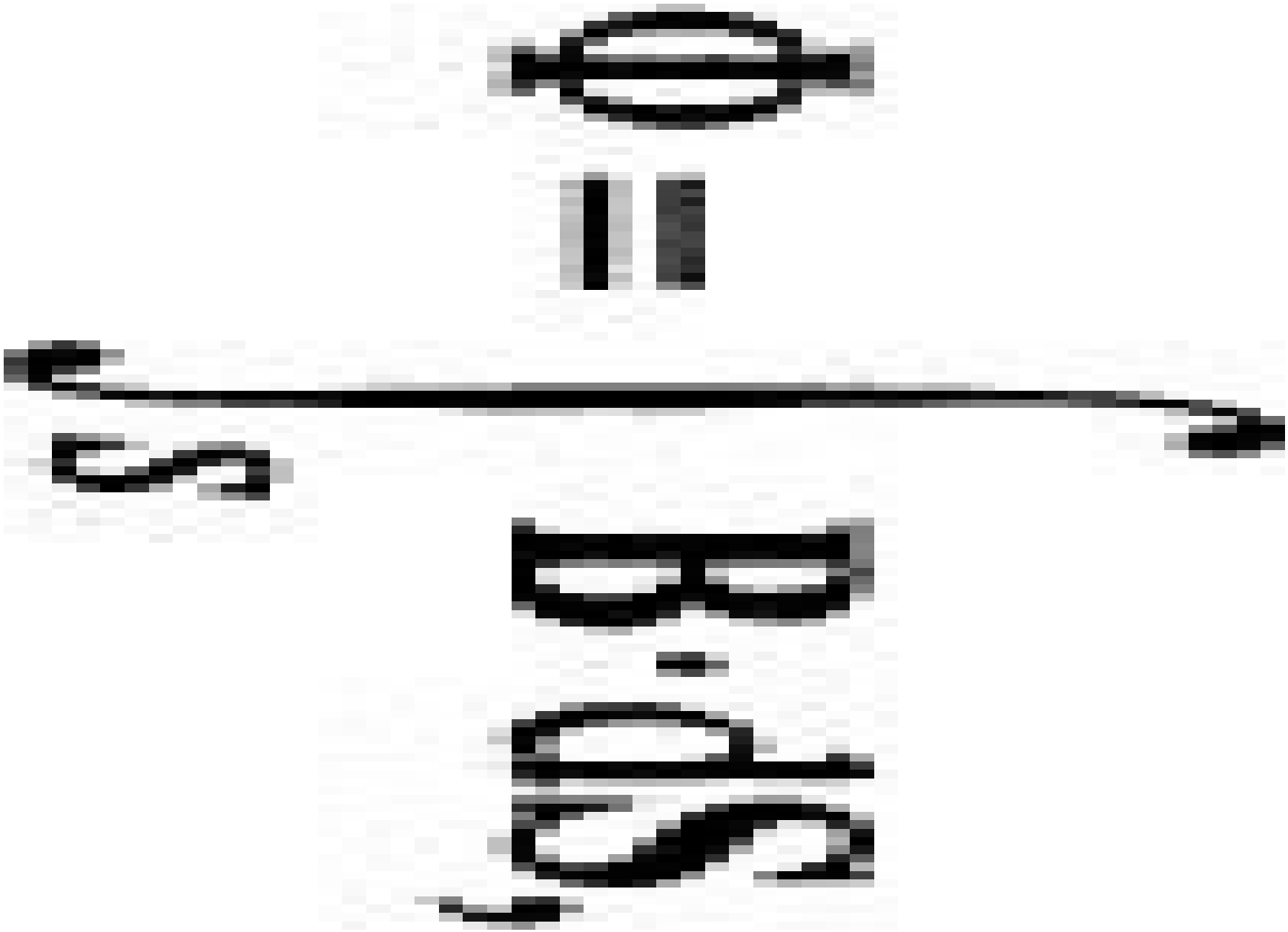
If the resistance of the coil is R the current is

(6.9)

Experiments show that the current is proportional to the rate of change with time of the total magnetic flux Φ through the circuit. Hence the size of the induced e.m.f. is proportional to the differential dΦ/dt,

(6.10)

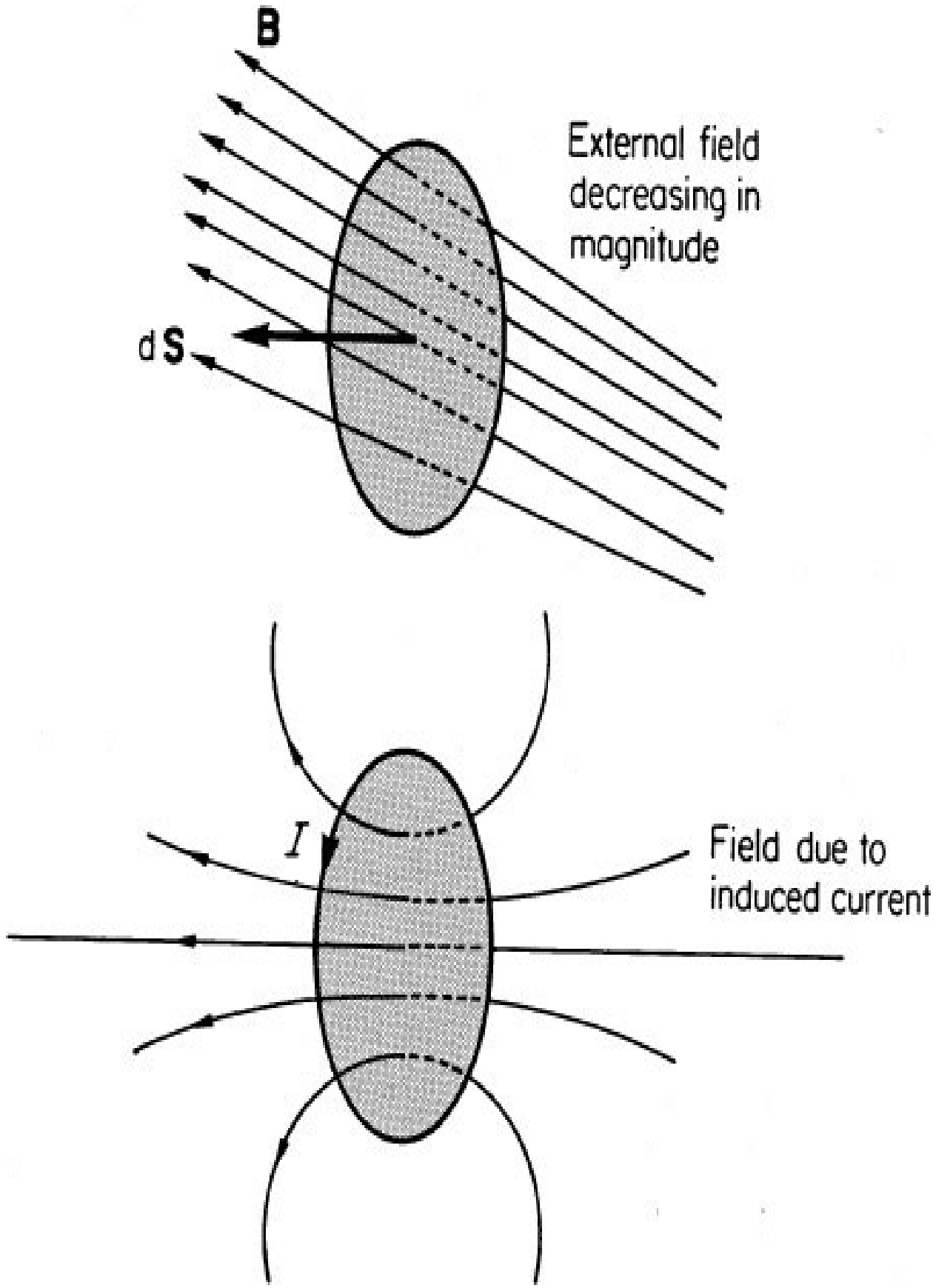
The direction of the current and e.m.f. is given by Lenz's law. This law says that the current induced is in a direction such that it produces a magnetic flux tending to oppose the original change of flux, i.e. tending to keep the total flux through the circuit constant. Lenz's law can be expressed mathematically. Figure 6.6 shows a coil in a magnetic field B. The magnetic flux through the coil is equal to



where for simplicity we choose S to be the plane surface whose rim is the coil, and the vectors dS to be in the direction which makes the flux positive. If the field is reduced the flux through the coil falls, and a current is induced which tends to stop the flux falling, i.e. it produces a field in the same direction as the field B. The induced current is thus in the direction shown in the lower drawing of Figure 6.6, and the line integral of the induced electric field is positive if we traverse the coil clockwise, looking from the right, in evaluating the integral. This sense of traversal of the coil is the one we must use since it is the one required by the right-hand screw rule together with the direction already chosen for the vectors dS. In our example the rate of change of flux $d\Phi/dt$ is negative, and so in order to obtain a positive e.m.f., clockwise around the coil, Equation (6.10) must be written

(6.11)

Figure 6.6. An illustration of Lenz's law. If the field B over the closed loop of wire (viewed from the left) decreases, a current I is induced in the loop. The direction of the current is such that the field to which it gives rise tends to stop the flux through the loop from falling, and so is in the direction shown in the lower figure.


B
External field
decreasing in
magnitude
dS
I
Field due to
induced current

If the quantities in Equation (6.11) are in SI units the constant of proportionality is found to be unity, as suggested by Equation (6.8), which applies to motional e.m.f.s.

All of the results of observations on induced e.m.f.s are summarized in Faraday's law, which says that the e.m.f. induced in a circuit is equal to the negative rate of change with time of the magnetic flux through the circuit.

(6.12)

or equivalently

(6.13)

where S is any surface whose rim is the loop under consideration. Faraday's law applies equally to any loop drawn in free space or in media, and not solely to a loop which is an actual conducting circuit, moving or otherwise. It is one of the fundamental laws of electromagnetism, on a par with Ampére's law and Gauss' law. The observation that changing magnetic fields produce electric fields was first made by Faraday around 1830, and independently by Henry and by Lenz near the same time.

6.1.3 Examples of induction

In this section we consider some examples of electromagnetic induction, and their analysis using Faraday's law, in order to become further acquainted with the use of the law and the ideas behind it.

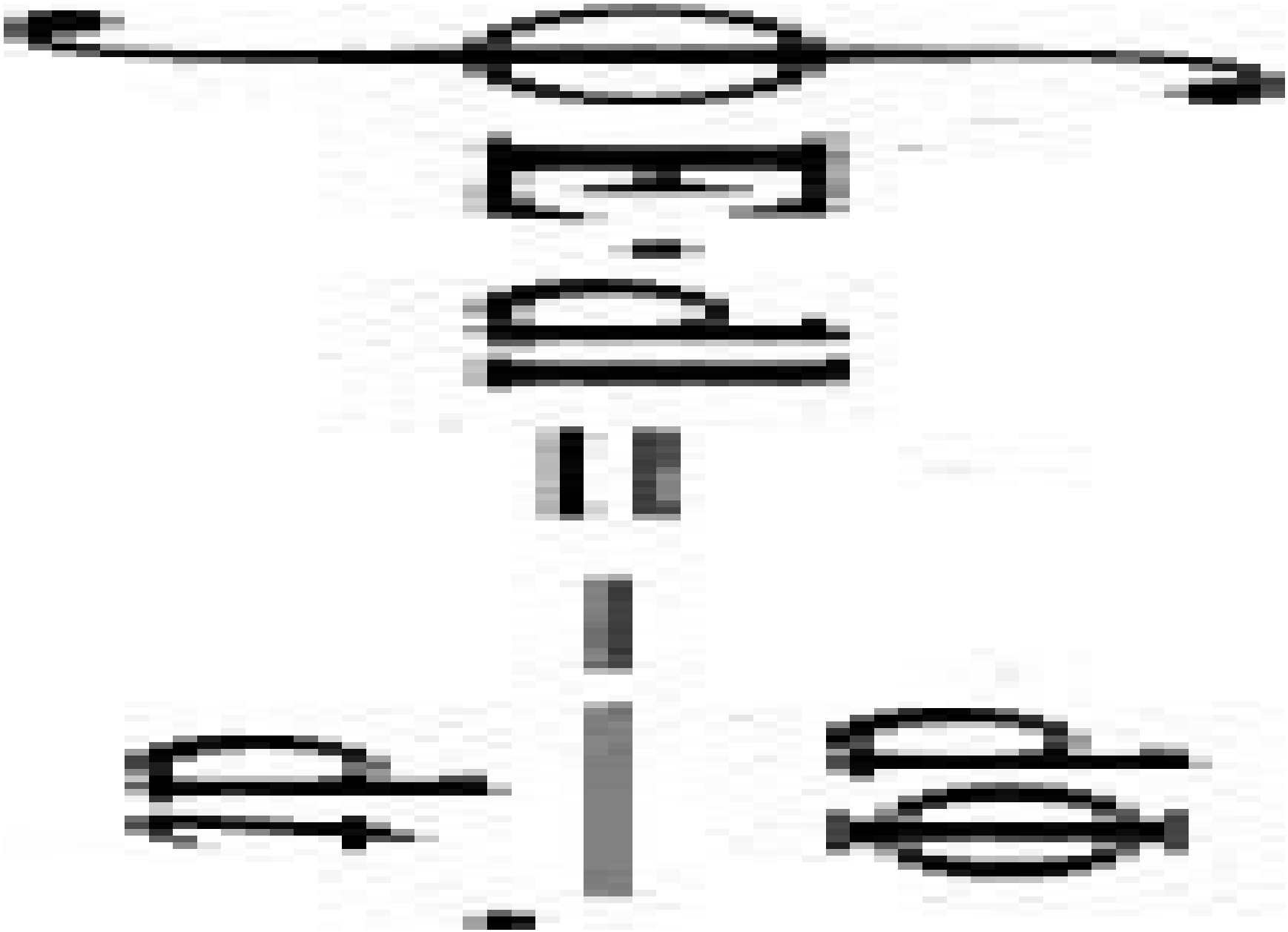
(i) A flat circular coil has 100 turns of wire each of radius 10 cm. A uniform magnetic field B exists in a direction perpendicular to the plane of the coil. The field is decreasing in strength at the rate of 10−2 Ts−1 The e.m.f. induced in the coil is equal to (number of turns) × (area) × (dB/dt) ≈ 0.03 V. If the resistance of the coil is 0.5 Ω, an induced current ≈ 0.06 A will flow. In this calculation we have neglected the flux through the coil due to its own field. This field is constant in this example, and so will have no effect. At the beginning and end of the change in the external field the current in the coil is not constant, and the resultant flux changes induce further e.m.f.s. This effect, which is called self-inductance, is discussed in section 6.2.

(ii) Faraday's law enables us to understand the origin of the changed energies of the electrons in atomic orbits when an external field is switched on. The electrons (e.g. in the orbits of Figures 5.2 and 5.3 used in the discussion of diamagnetism) are either speeded up or slowed down by the electric field induced when the external field is switched on.

Consider the orbit shown in Figure 5.2. The area of the orbit is πr^2. The magnetic flux Φ through this area increases from zero* to the value Φf during the time it takes to establish the external field B,

(6.14)

While the field is changing there is an e.m.f. in the orbit given by



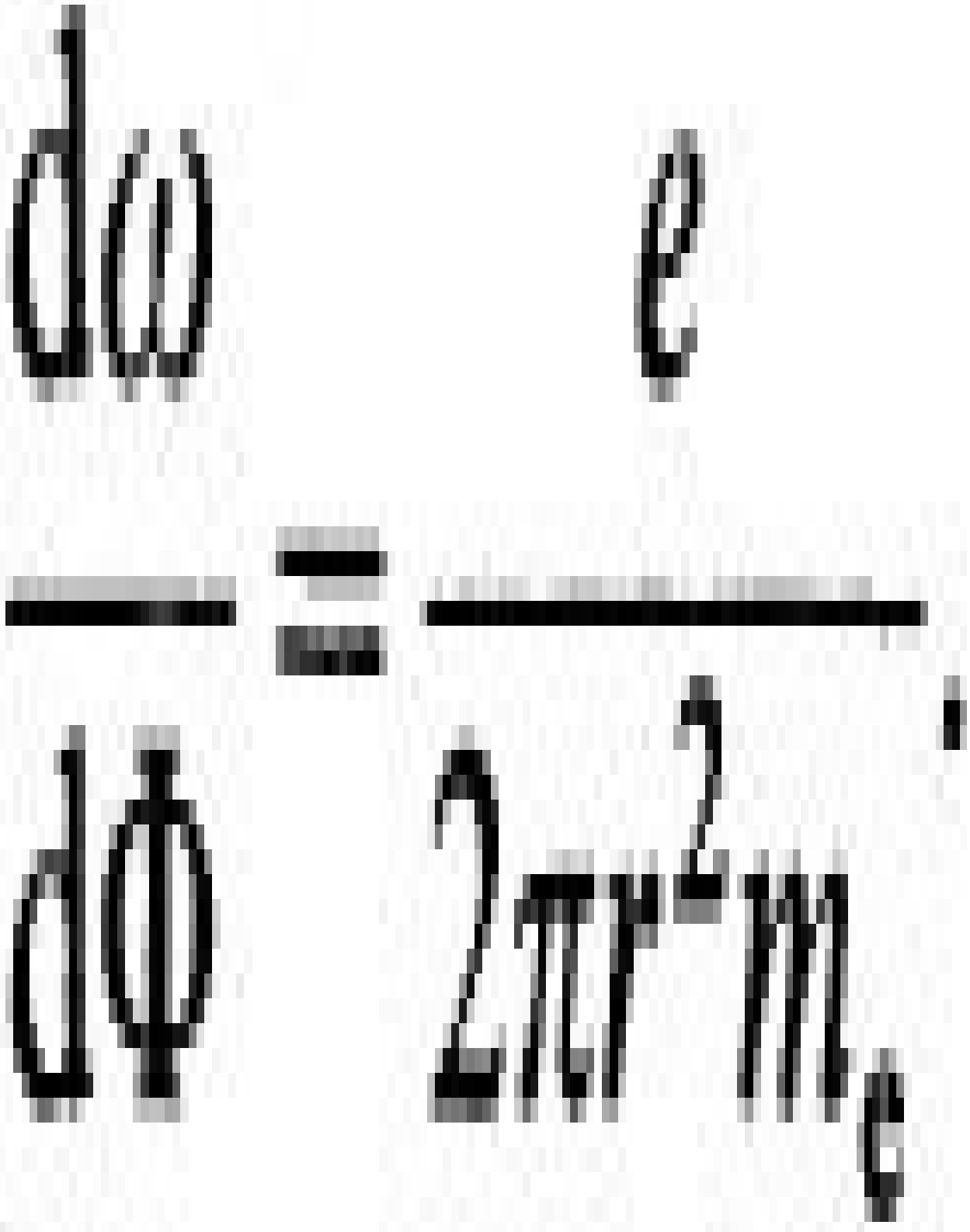
The e.m.f. acts in a direction opposite to the electron's motion and so gives a force on the electron which speeds it up. Its angular frequency increases from $\omega 0$ to the final value ωf. At an intermediate time t, when the angular frequency is ω the electron makes a fraction $\omega dt/2\pi$ of a complete turn in a time dt. The work done on the electron during this small time by the induced electric field is

$$dW = \frac{\omega}{2\pi} dt \cdot e \frac{d\Phi}{dt}$$

$$= \frac{e\omega}{2\pi} d\Phi.$$

This work is equal to the increase in kinetic energy of the electron, assuming that the electron’s orbit remains unchanged, and hence

from which



By integrating this equation we determine the change in angular frequency,

$$\omega_f - \omega_0 = \frac{e\Phi_f}{2\pi r^2 m_e}$$

$$= \frac{eB}{2m_e},$$

from Equation (6.14). The change in angular frequency is equal to the Larmor angular frequency ωL. The result is the same as obtained in section 5.1.1.

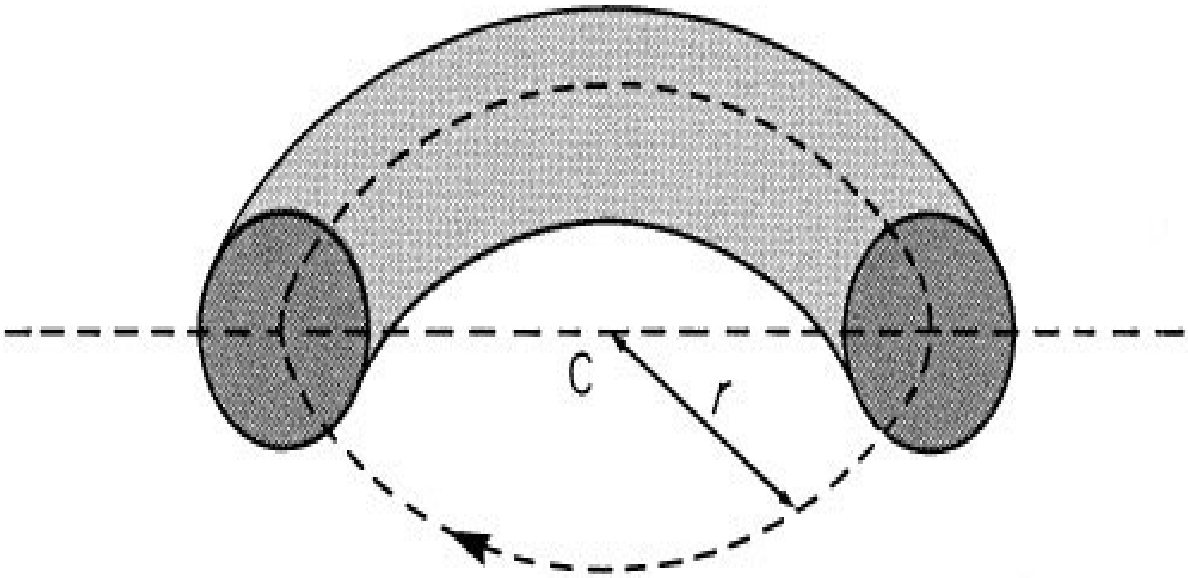
(iii) The betatron, which is a machine designed to accelerate electrons to energies of about 50MeV, gives a practical demonstration that Faraday's law holds for the electric field around a closed loop in free space, and not only for loops within conducting materials. Betatrons were used in hospitals for radiation therapy. Although they have nowadays been replaced by other devices they provide an instructive example of how electric and magnetic fields can be combined to accelerate electrons in a controlled way. In the betatron a changing magnetic field is used to accelerate electrons in a circular orbit in an evacuated chamber. The field is designed so that the electrons, as they are accelerated, remain in the same orbit.

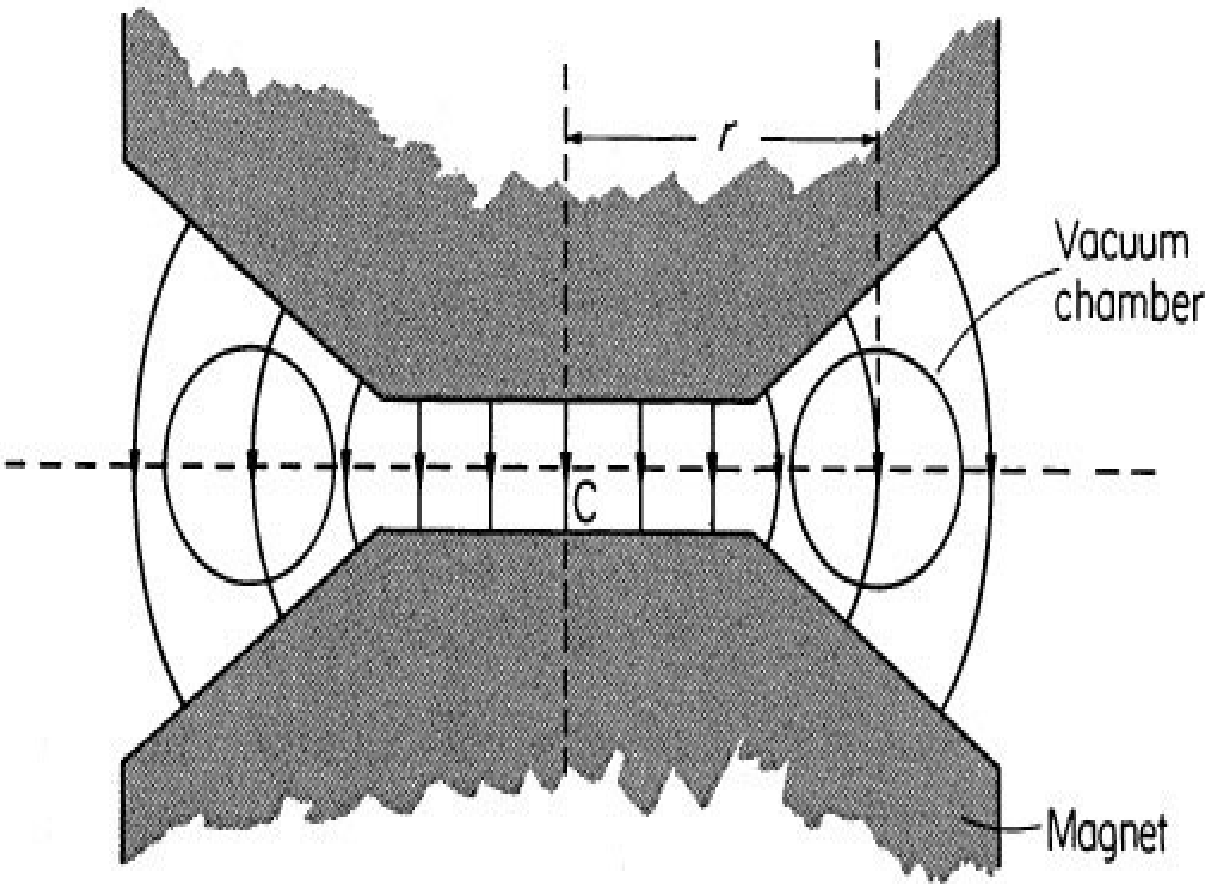
Consider an electron moving in an orbit of radius r in a vacuum chamber situated between the poles of an electromagnet, as shown in Figure 6.7. If the magnetic field increases an e.m.f. is induced in the orbit. Since there is symmetry about the axis of the magnet, the electric field induced in the magnet gap depends only on the perpendicular distance from the axis. For the same reason the induced field at any point is in a direction tangential to a circle drawn around the axis and passing through the point. The electric field induced in the electron's orbit therefore has the same magnitude E at all points in the orbit,

(6.15)

Figure 6.7. Aschematic representation of a betatron. The top drawing shows the vacuum chamber situated in a magnetic field which decreases with distance from the centre C. If the field is in the direction shown and increases in magnitude with time the electrons are accelerated in a clockwise direction, looking downwards. The bottom drawing shows a section through the electromagnet and

vacuum chamber.


C
r

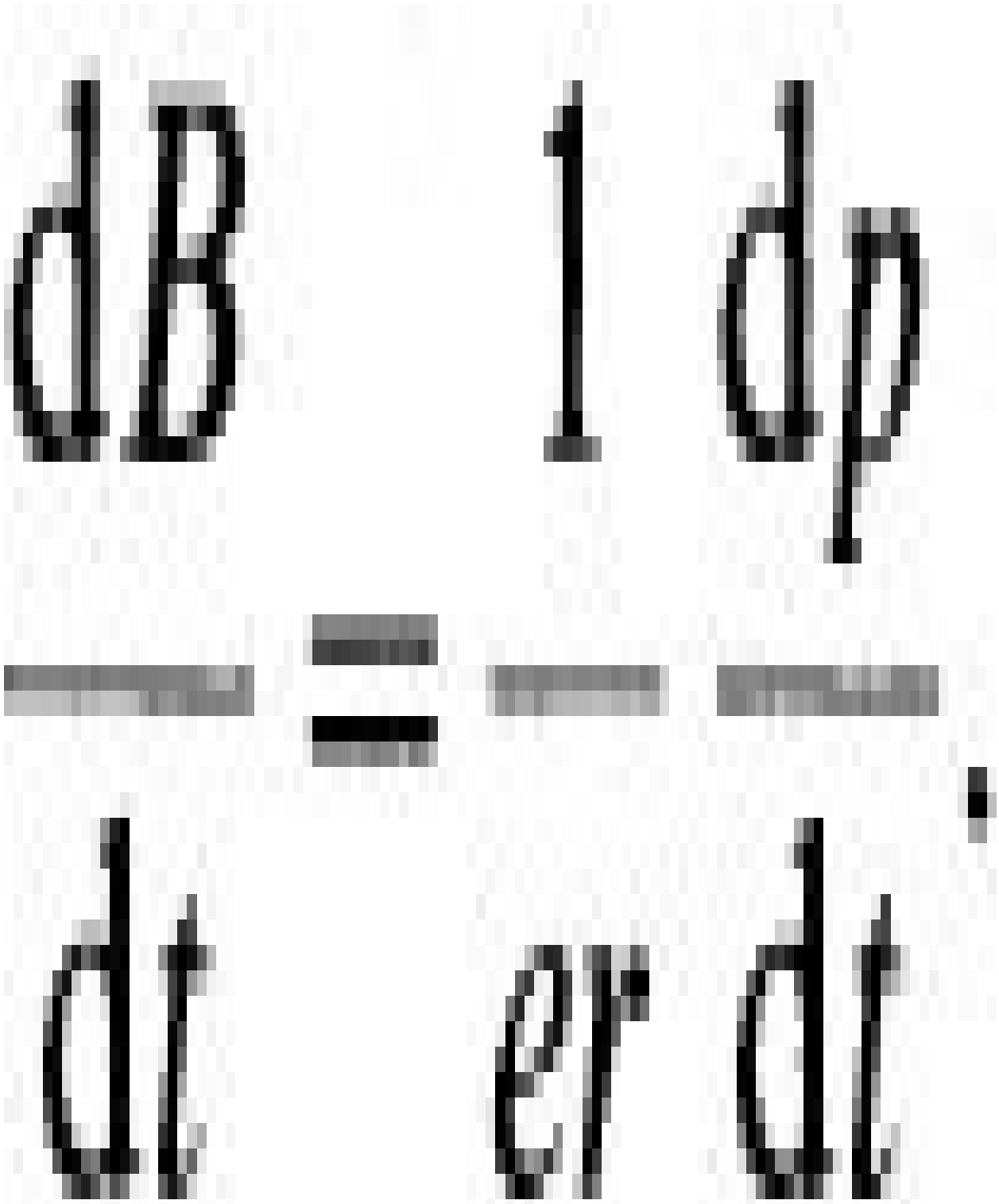

r
Vacuum chamber
C
Magnet

where Φ is the magnetic flux through the orbit. The direction of the induced field in Figure 6.7 is anticlockwise viewed from above. This field produces a force on the electron accelerating it in a clockwise direction, as indicated by the arrow on Figure 6.7.

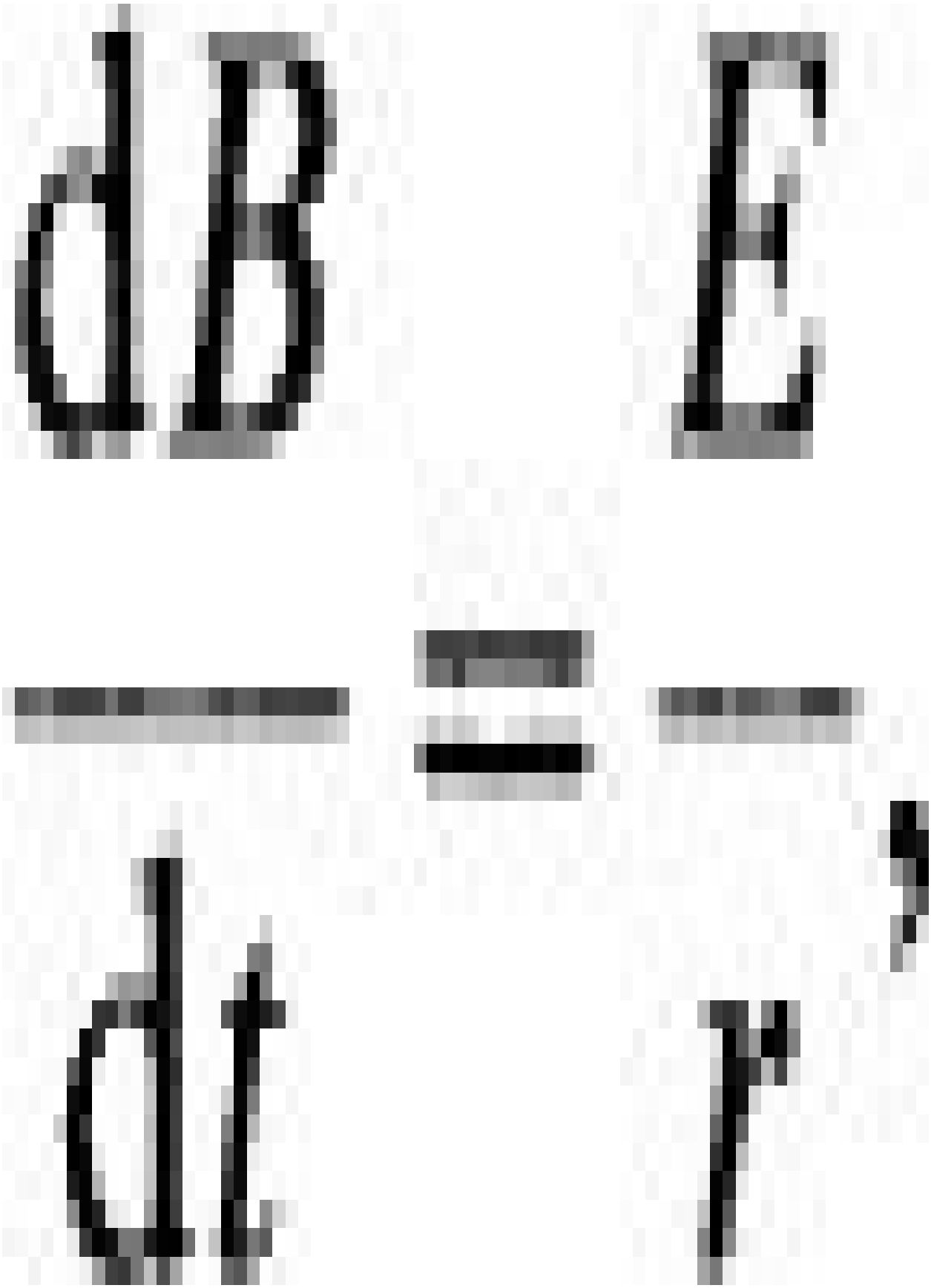
If the electron is to stay in its orbit the field B over the circle of radius r must at all times satisfy Equation (4.58),

(6.16)

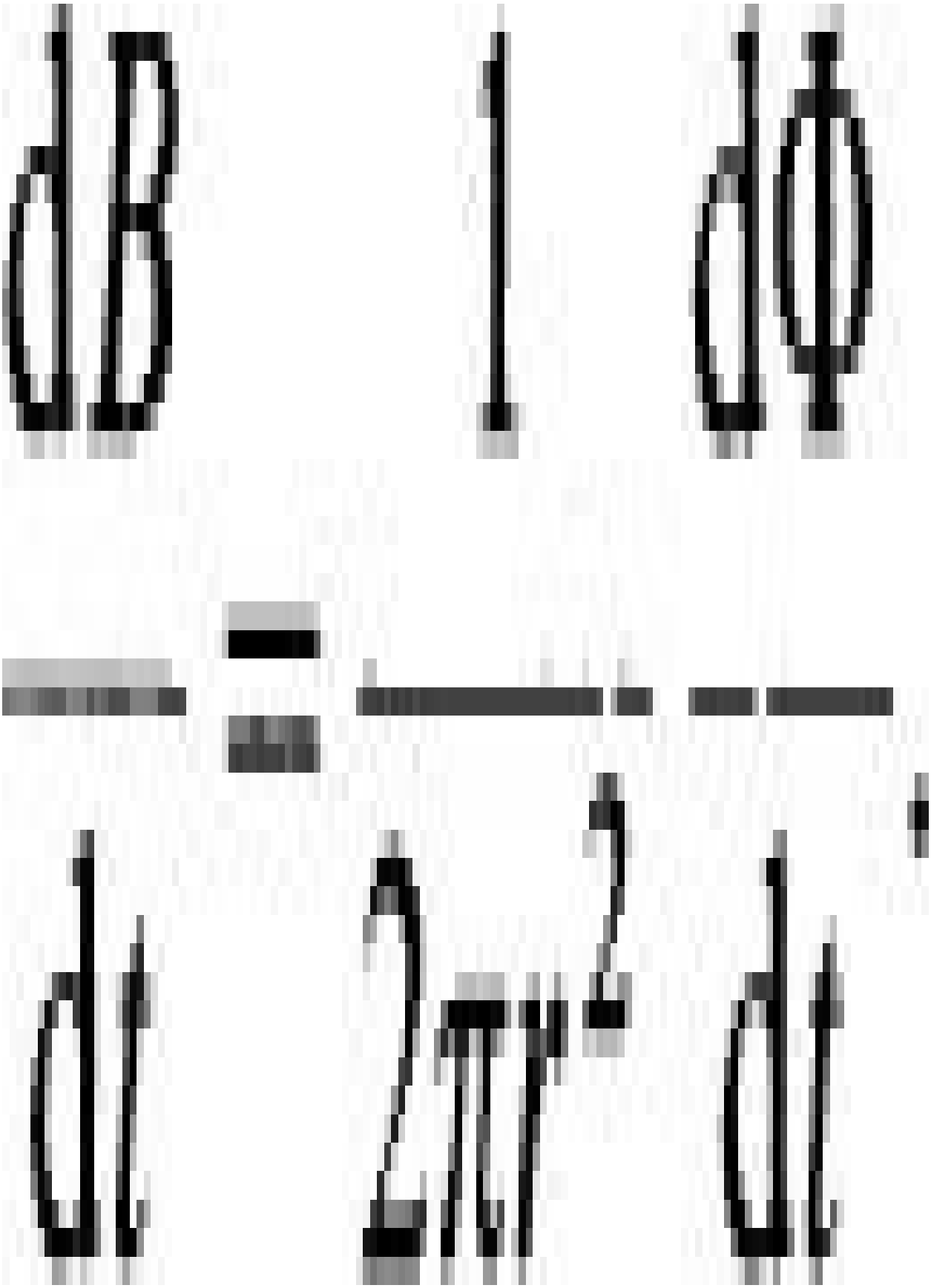
where p is the momentum of the electron. Differentiating this equation with respect to time



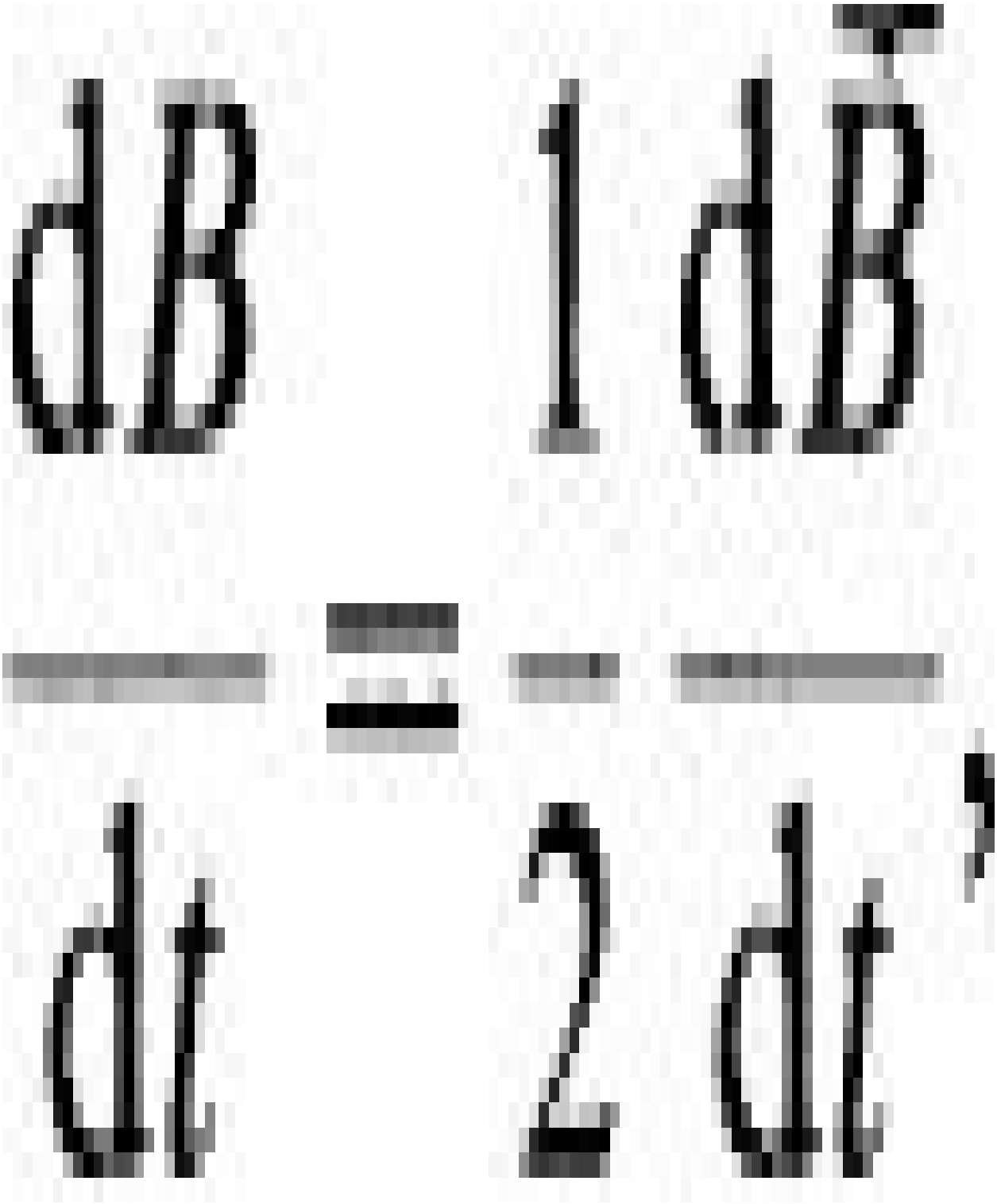
Since the rate of change of the electron’s momentum is equal to the force eE,



and substituting for the field E from Equation (6.15)



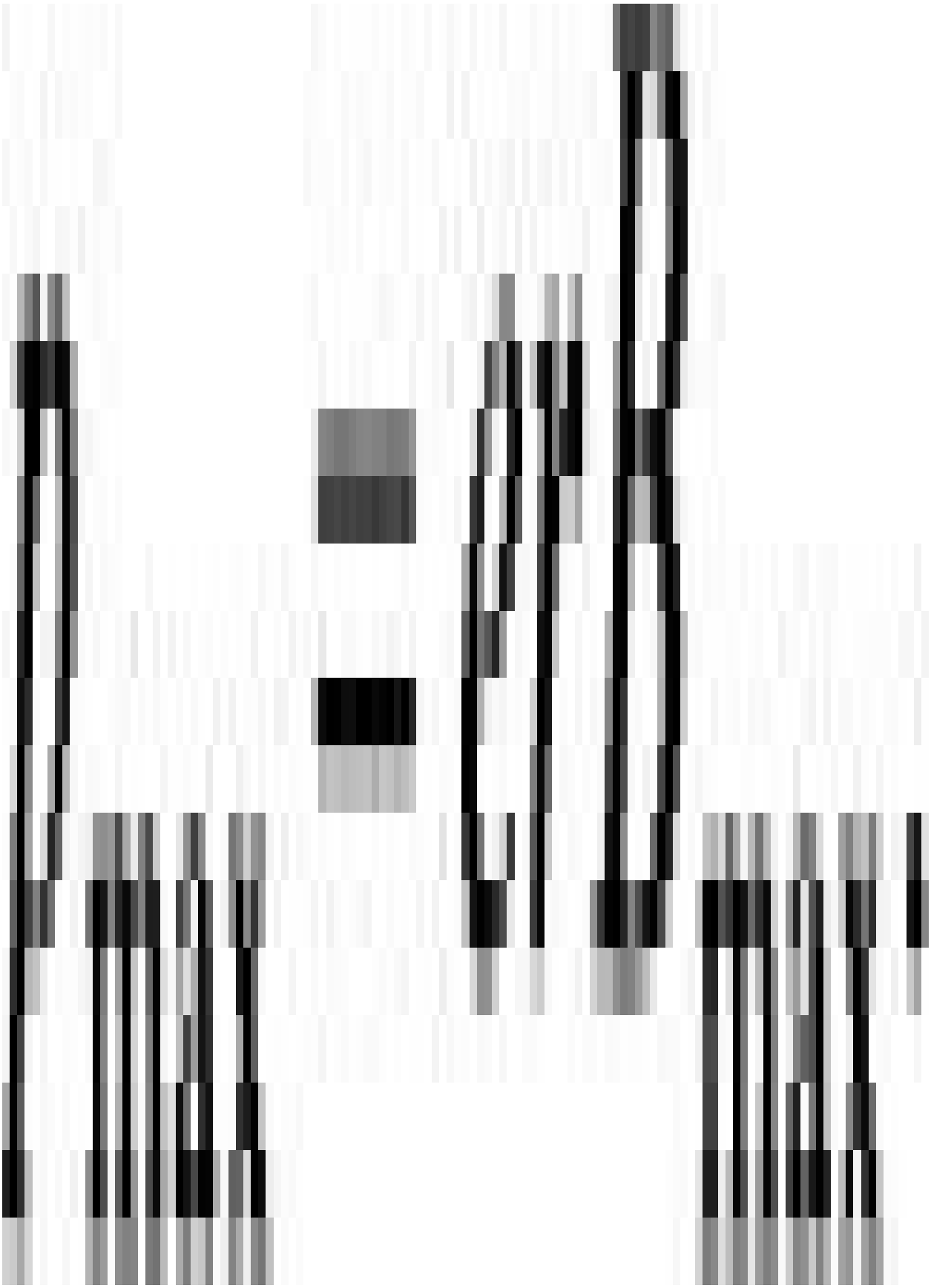
The flux Φ through the orbit is equal to πr^2 , where is the average field over the area πr^2 of the orbit. Hence



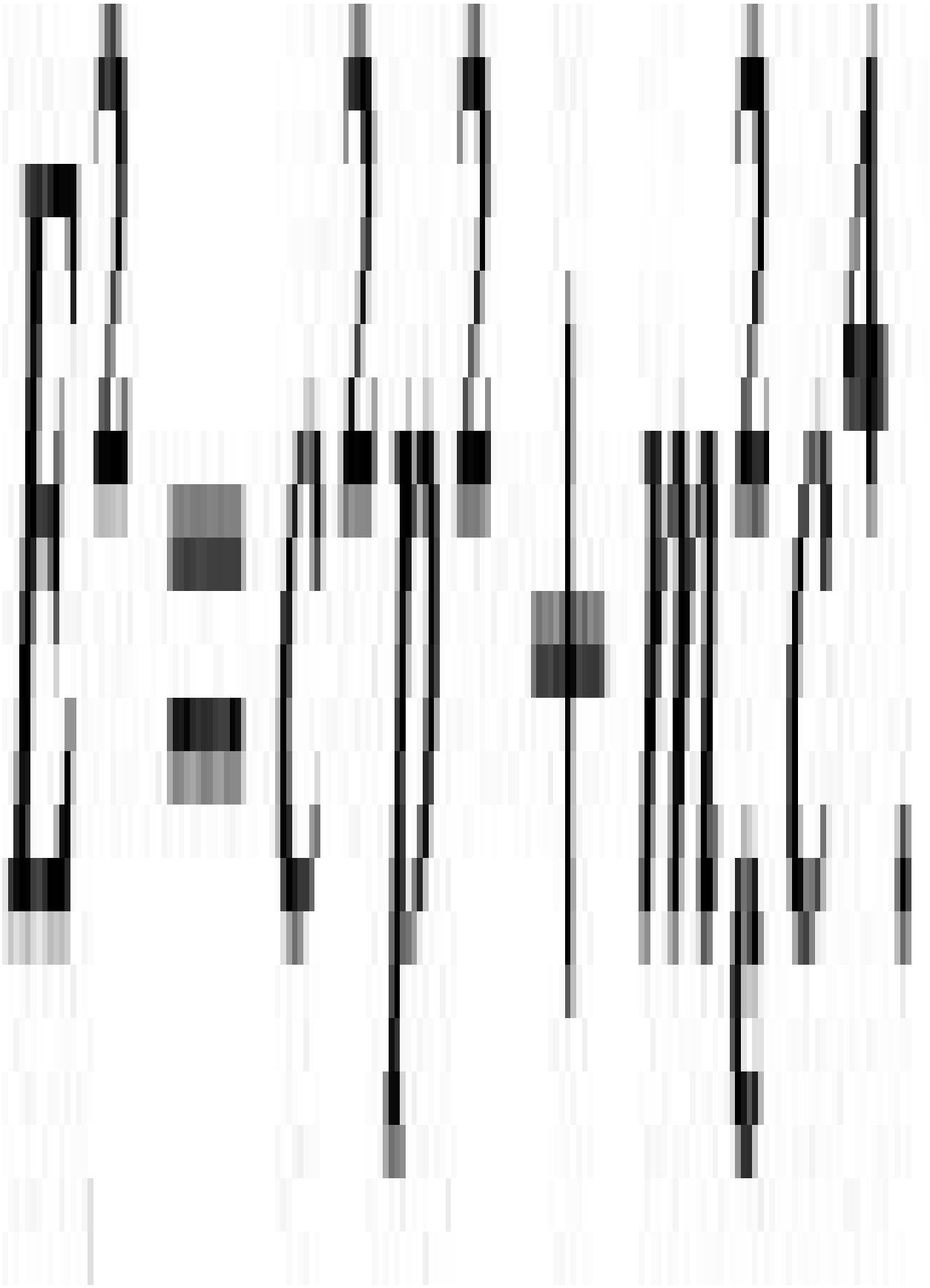
and for the electron to stay in the same orbit this equation must be satisfied at all times during the acceleration. The condition for a fixed orbit is thus that

and this is arranged by making the magnetic field decrease outwards from the middle by increasing the gap between the pole faces, as shown in the figure.

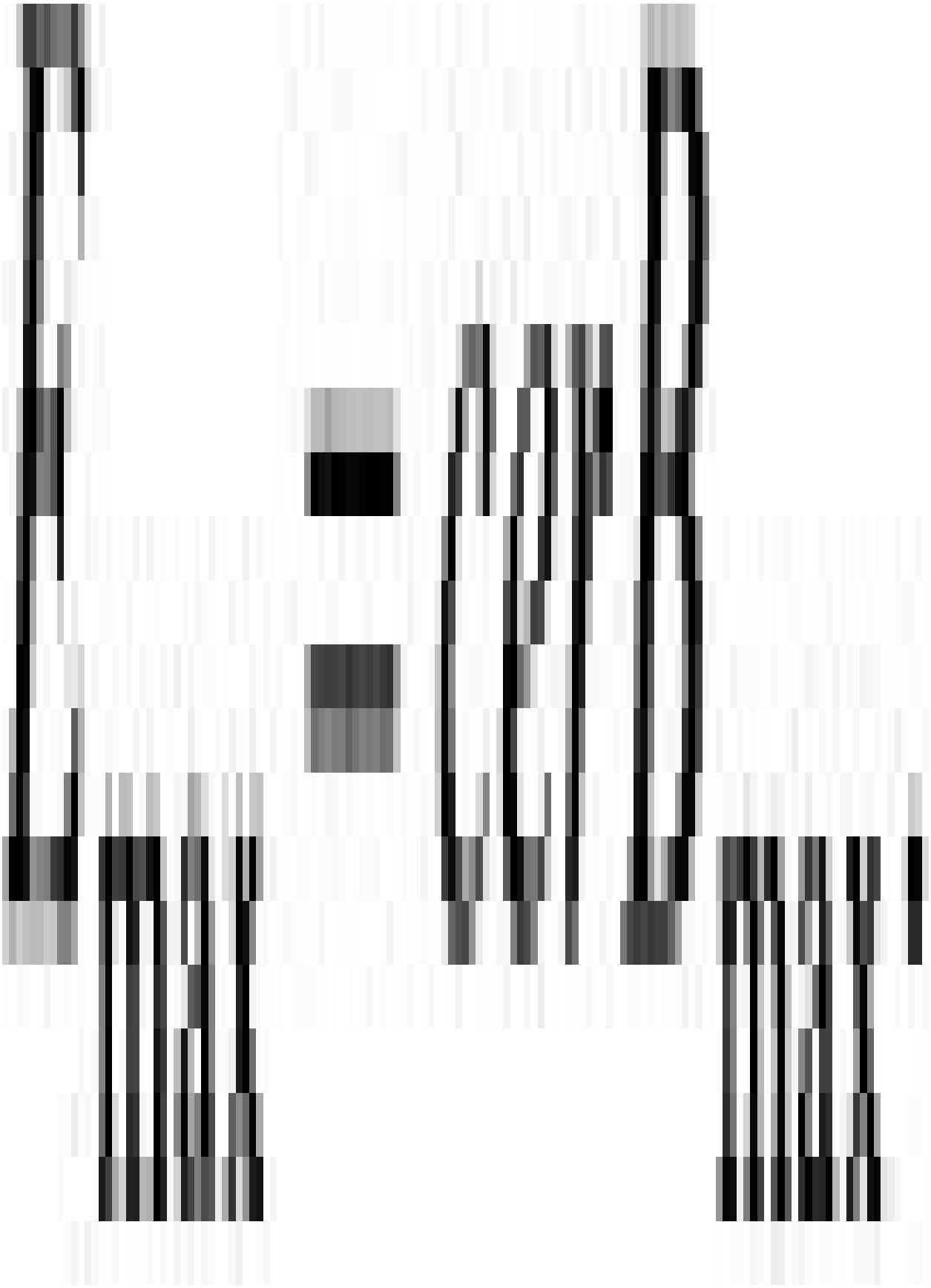
If the maximum magnetic field that can be attained at the orbit of the electron is Bmax, Equation (6.16) enables us to calculate the maximum momentum acquired by the electrons,



The electrons in the betatron are moving at speeds comparable with the speed of light in free space, c, and the calculation of their energy has to be done relativistically. The formulae used so far are valid in relativistic dynamics. The relation between the total energy E of the electrons, their rest mass me and momentum p is



In this example $c^2p^2 \gg c^4$ and hence $E \approx cp$; $E_{max} \approx cp_{max}$. The maximum energy acquired by the electrons E_{max} is thus given by



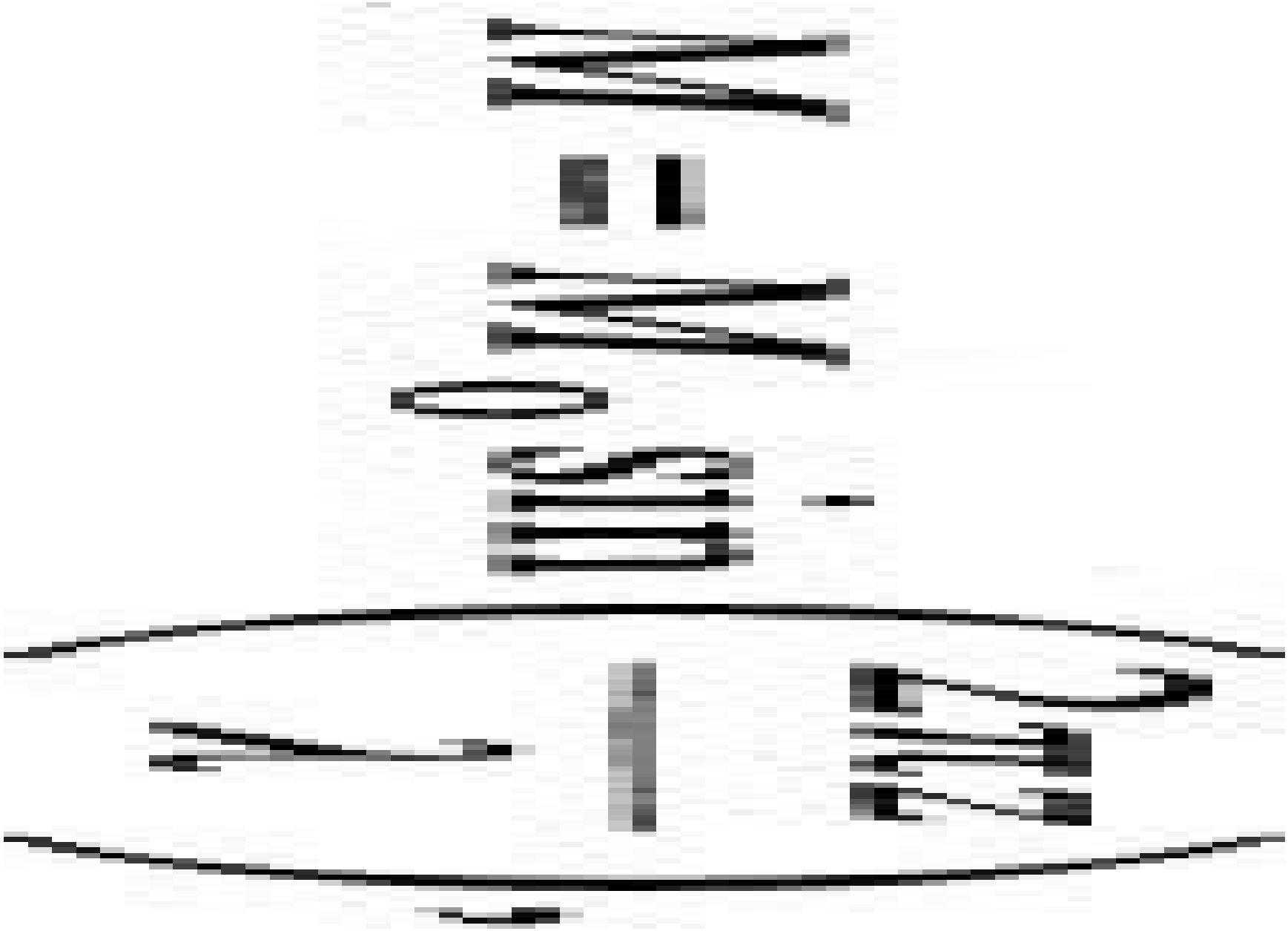
For Bmax = 0.25 T, and r = 0.5 m the electrons will be accelerated to energies of about 37 MeV.

(iv) Our final example deals with the retrieval of information stored on magnetic tape. Tape storage and retrieval play a very large part in modern life, from producing music to accessing complicated scientific or engineering data. A magnetic tape consists of a thin ribbon of plastic on whose surface has been coated a very thin layer of magnetizable substance. This layer is a dispersion of small particles of hard magnetic material in a binding agent. Hard magnetic materials were discussed at the end of section 5.4.1 and have magnetizaton curves similar in general to that shown in Figure 5.18. Once magnetized in a certain direction by an external field they require a large field in the opposite direction to flip the magnetization around.

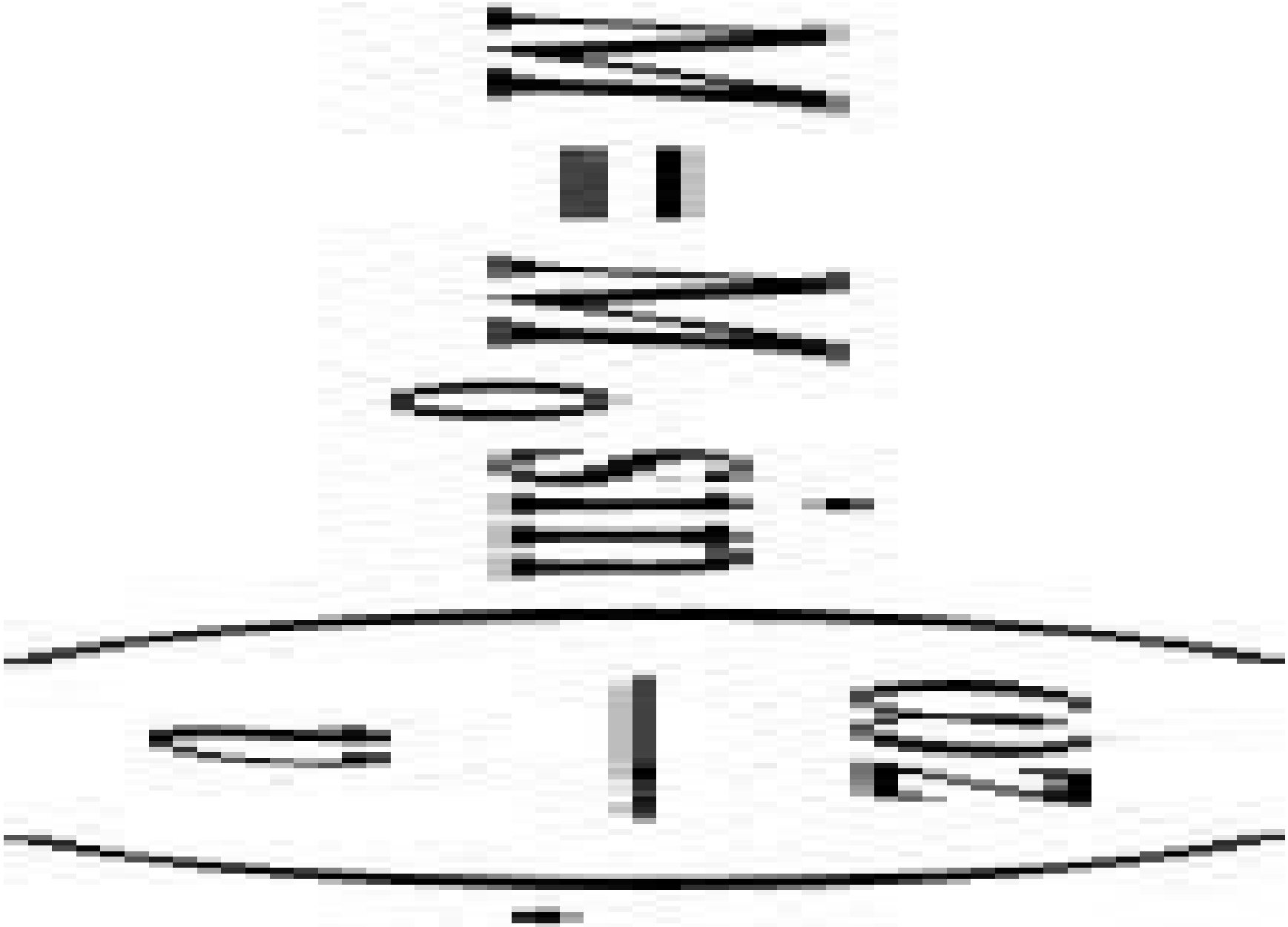
The way information is stored on magnetic tape is by ordering the magnetization of the magnetic particles as the tape is passed at constant speed through a recording head. This head consists of a small electromagnet, made of high permeability, soft material, in which the field in the gap is controlled by the current in the magnet coils, thus converting electrical current signals into magnetic fields. The tape passes over the top of the magnet gap and as it leaves the gap is magnetized by an amount depending on the field there at the time. This process converts the electrical signals, which contain the information to be stored, into a magnetization distribution along the length of the tape. For retrieval the tape is passed at the same speed as it was recorded over an electromagnet similar to the one used in the recording head and the process is reversed: the flux variations in the magnet gap caused by the magnetic fingerprints on the tape are now converted to current signals in the magnet coils. These can be decoded to give the original current signals which impressed the magnetization on the tape.

Suppose we record a current signal, of a single frequency f and constant amplitude, on a tape moving at constant speed v over a recording head. To a

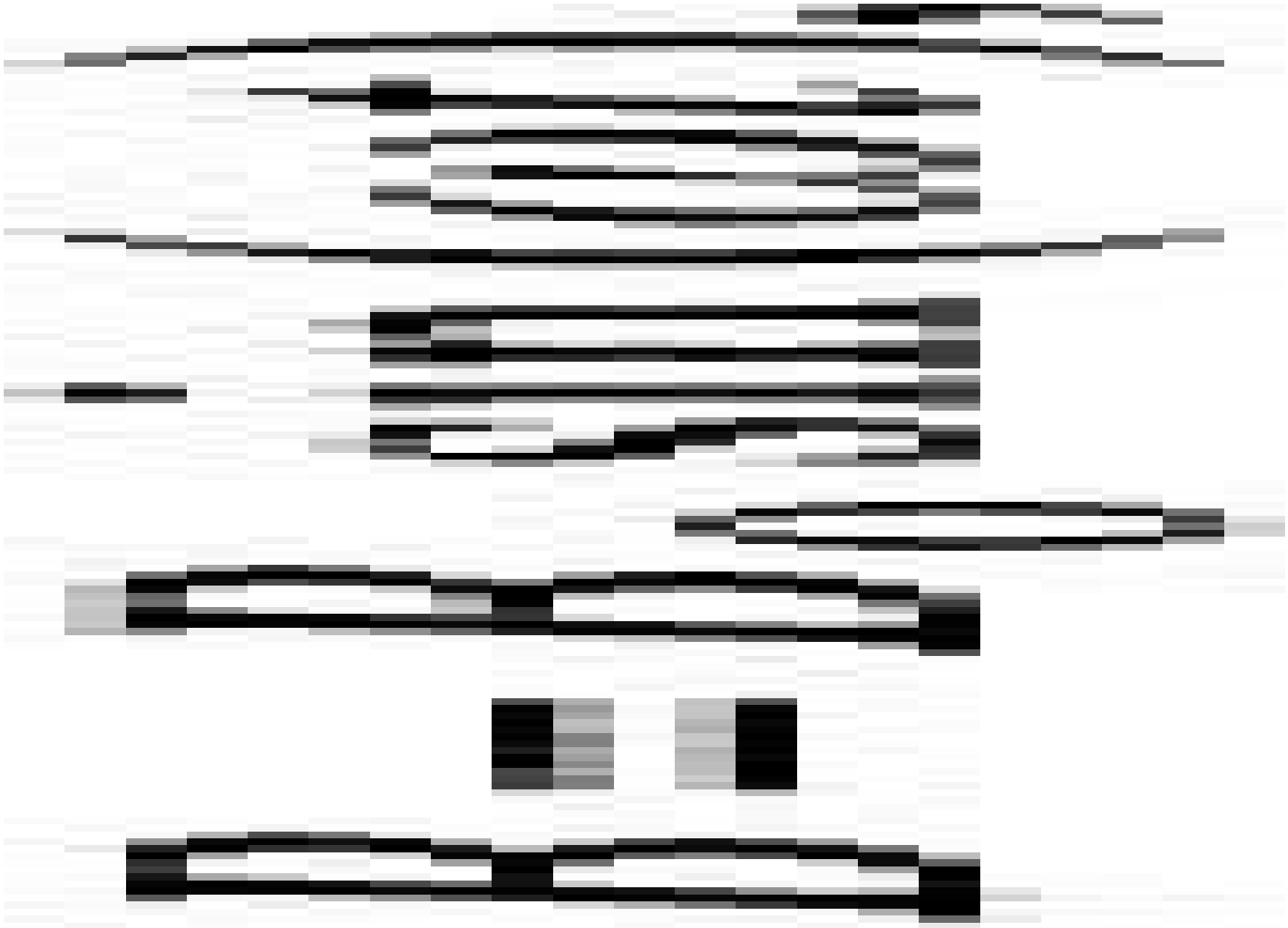
good approximation the field in the gap and the associated magnetization on the tape are proportional to the current in the coils. The magnetization distribution on the tape then varies sinusoidally with distance z along the tape and the size of the magnetization in the direction in which the tape moves can be written

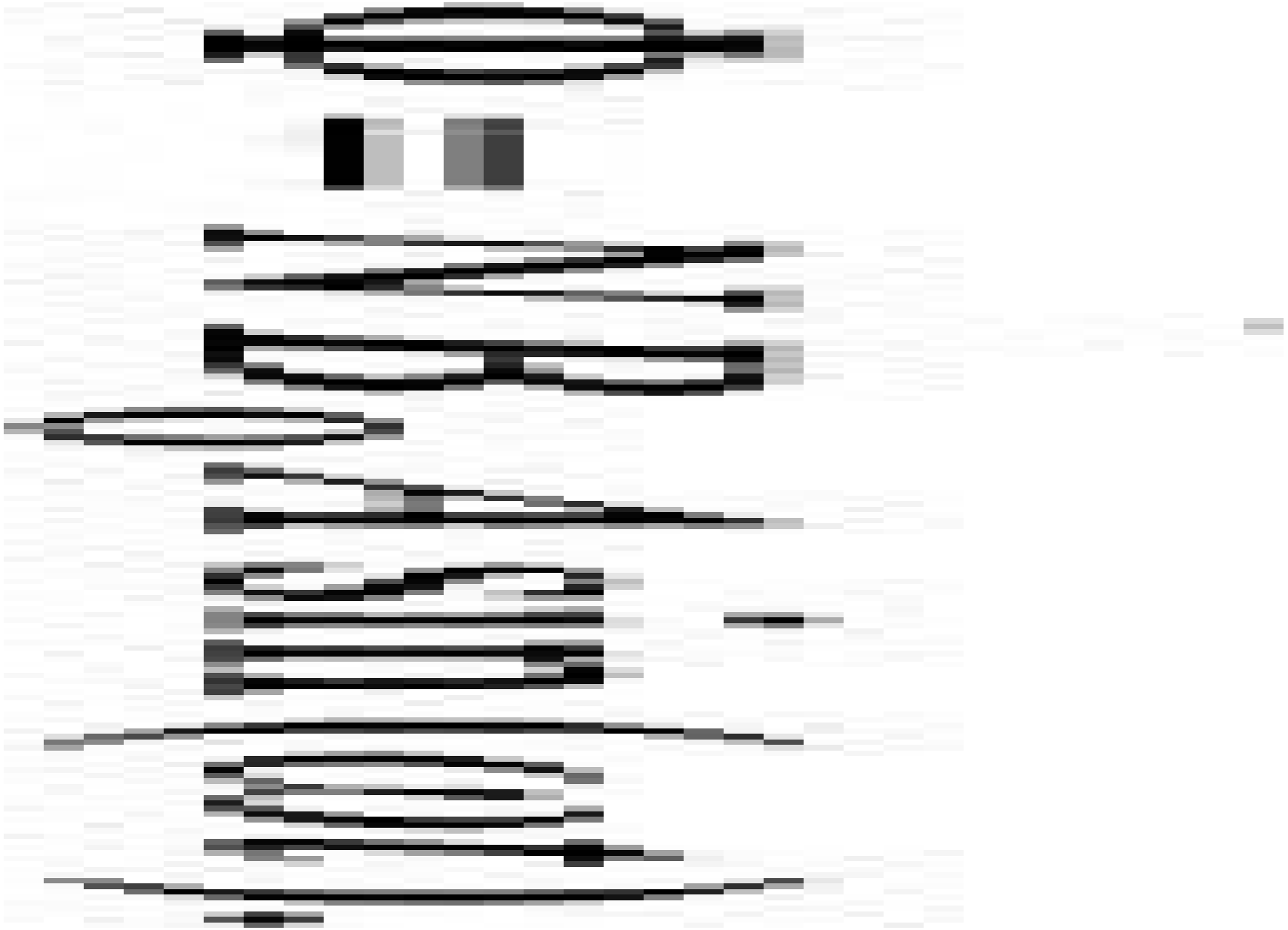


where $\lambda = v/f = 2\pi/\omega$. Thus

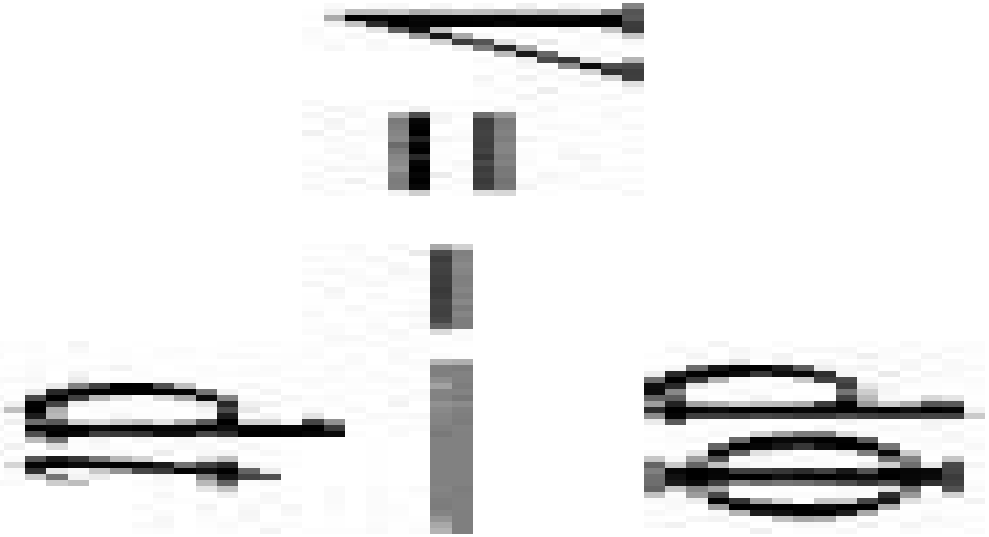


When the tape is played back at the same speed past the narrow gap in the magnet of a reproducing head the magnetization on the tape produces a field in the magnet gap which varies sinusoidally with time at angular frequency ω. This field in turn is associated with a field of magnitude B in the magnet core which, again to a good approximation, is proportional to the magnetization on the tape. If there are N turns on the coils wound round the core and the core has cross-sectional area A there is a flux linkage Φ with the coils and we have

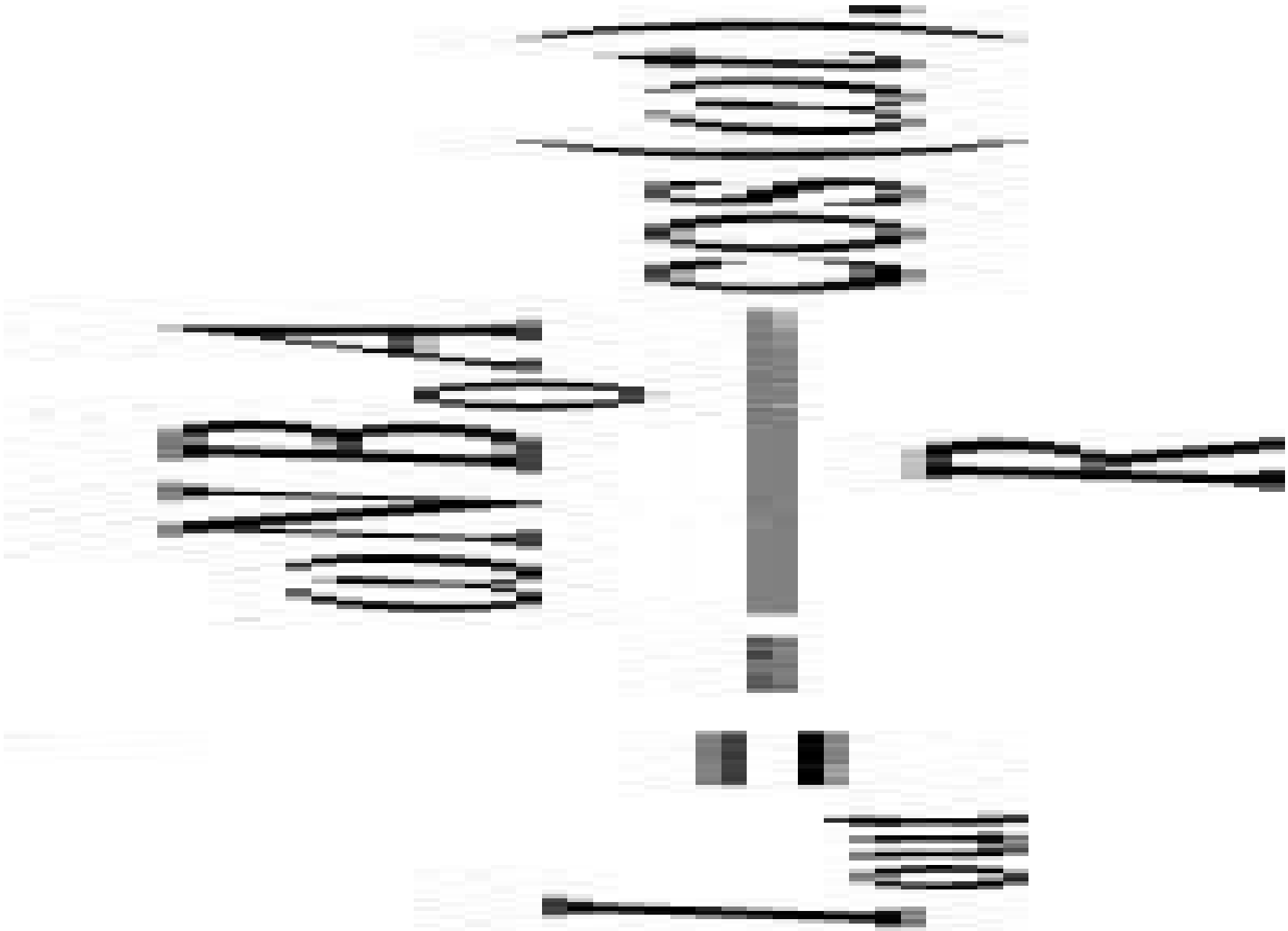




The changing flux linkage as the tape moves induces an e.m.f. V in the coils, and from Faraday’s law



If the resistance of the circuit through which this e.m.f. passes a current is R the output current Iout is given by



We see at once that the size of the output current, because of the electromagnetic induction process, is proportional to the time derivative of the original recorded current and not to that current itself. However the original current form can be regained by using an integrating electronic circuit to process the output current. This process of recording and reproduction is shown schematically in Figure 6.8.

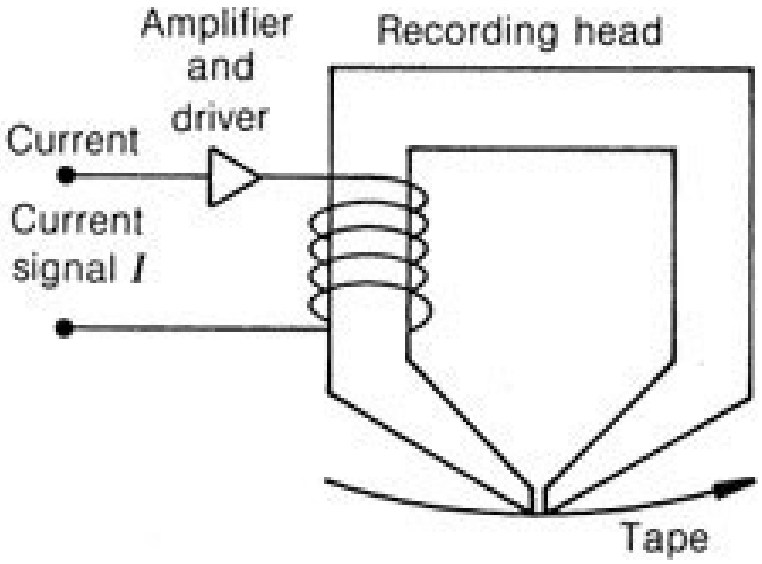
In a typical reproducing head the number N is about 200, the area A about $4 \times 10^{-7} m^2$, and the field amplitude B0 about 10^{-3} T. At a frequency of 10^3 Hz ($\omega = 2\pi \times 10^3 s^{-1}$) such a head produces a voltage amplitude of about 0.5 mV.

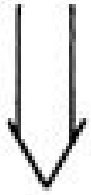
Of course one never stores a simple signal such as a pure sinusoid on tape. Information to be stored is in the form of complicated functions of time. In the case of music recording, for example, the currents are outputs from microphone systems and are related to the sound waves which cause the motion of the diaphragm in the microphone. However, any function of time can be expressed in terms of its component frequencies using Fourier analysis and the above simplified outline using a pure sine wave illustrates the principles on which tape storage and retrieval are based. Nowadays more and more tape storage is in the form of digital information. Analogue signals are digitized and stored on tape as magnetizations of fixed size. This greatly reduces problems such as those associated with noise and variable tape speed and makes data storage and retrieval highly reliable.

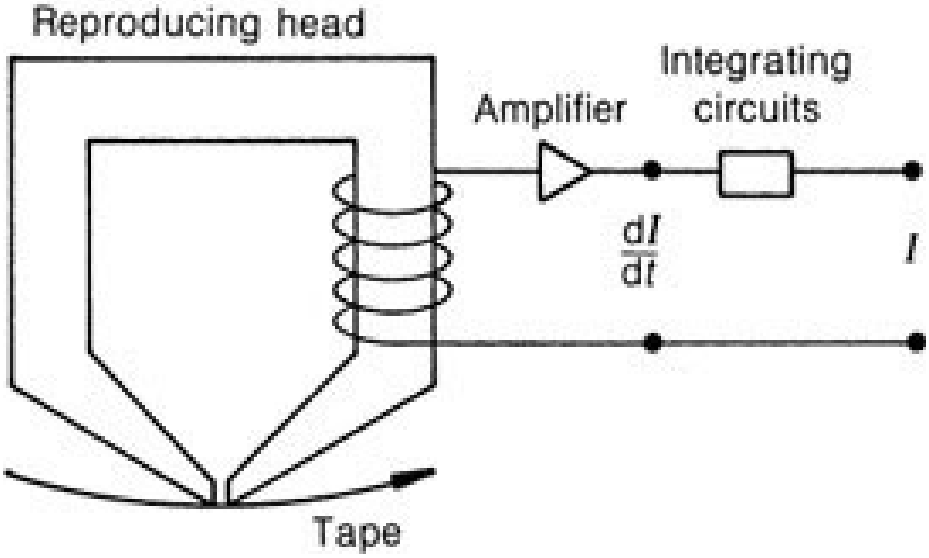
There are many other practical applications and results of electromagnetic induction. One of the most important applications is the generation of electrical power in power stations. Some kinds of electric motor depend on induction for their operation. Induced e.m.f.’s are also used to couple together different parts of a circuit as in power transformers. Induced currents are not always welcome. For example, electric fields are induced in the metal parts of electrical machinery as external magnetic fields change. The electric fields produce currents in the metal conductors, called eddy currents, and these cause unwanted energy losses and other effects. In the iron cores of transformers or electromagnets,

precautions are often taken to minimize the effects of eddy currents by splitting the iron into thin layers insulated from each other so that eddy currents can not pass between them. The generation of electric power and transformers are discussed in later chapters.

Figure 6.8. A schematic picture of the way information is written on to magnetic tape and read from it. An electromagnet is used to magnetize the tape, and electromagnetic induction with a second magnet to interrogate and interpret the magnetization.


Amplifier
and
driver
Recording head
Current
Current
signal I
Tape

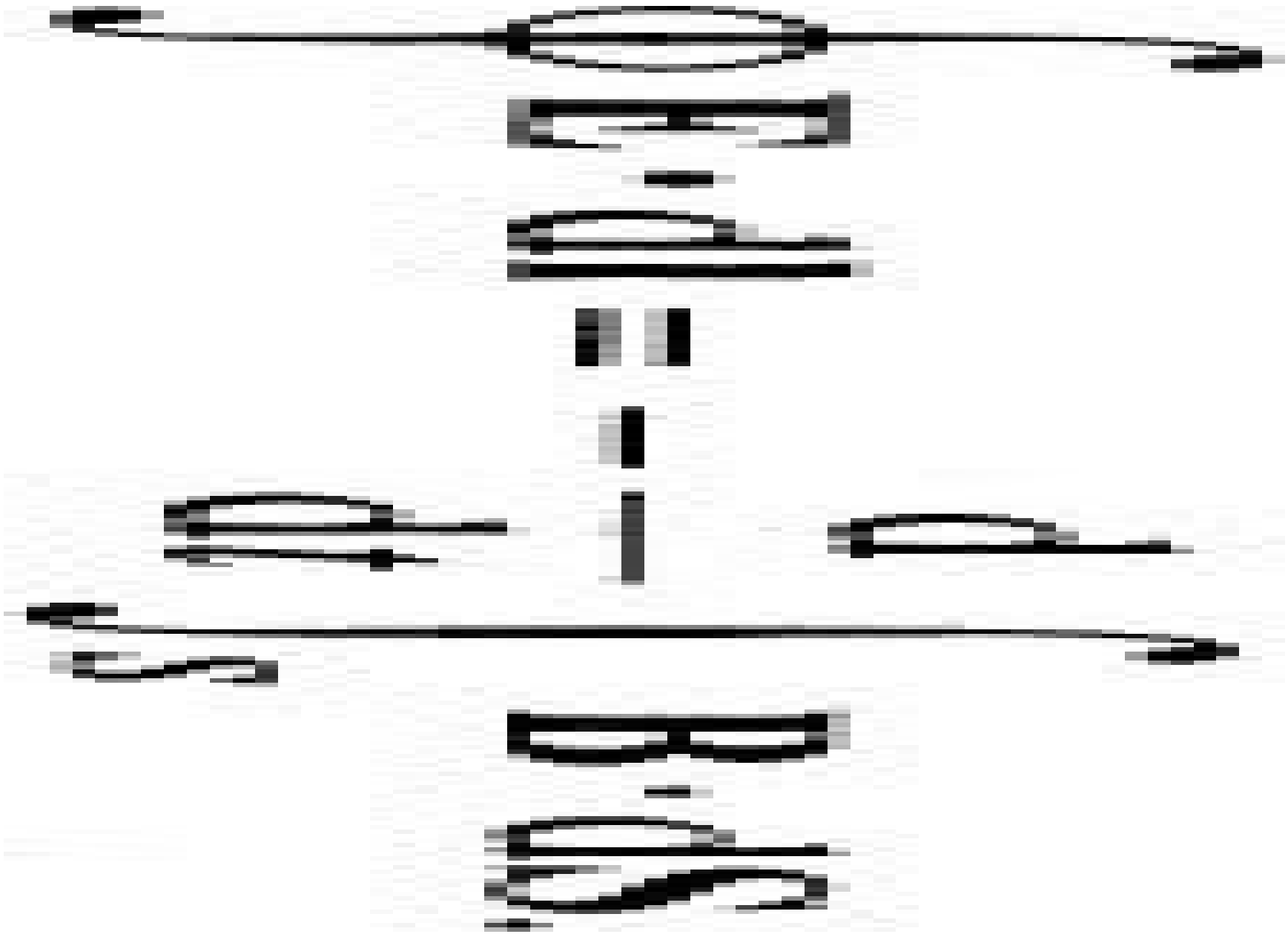



Reproducing head
Amplifier
Integrating
circuits
dI/dt
I
Tape

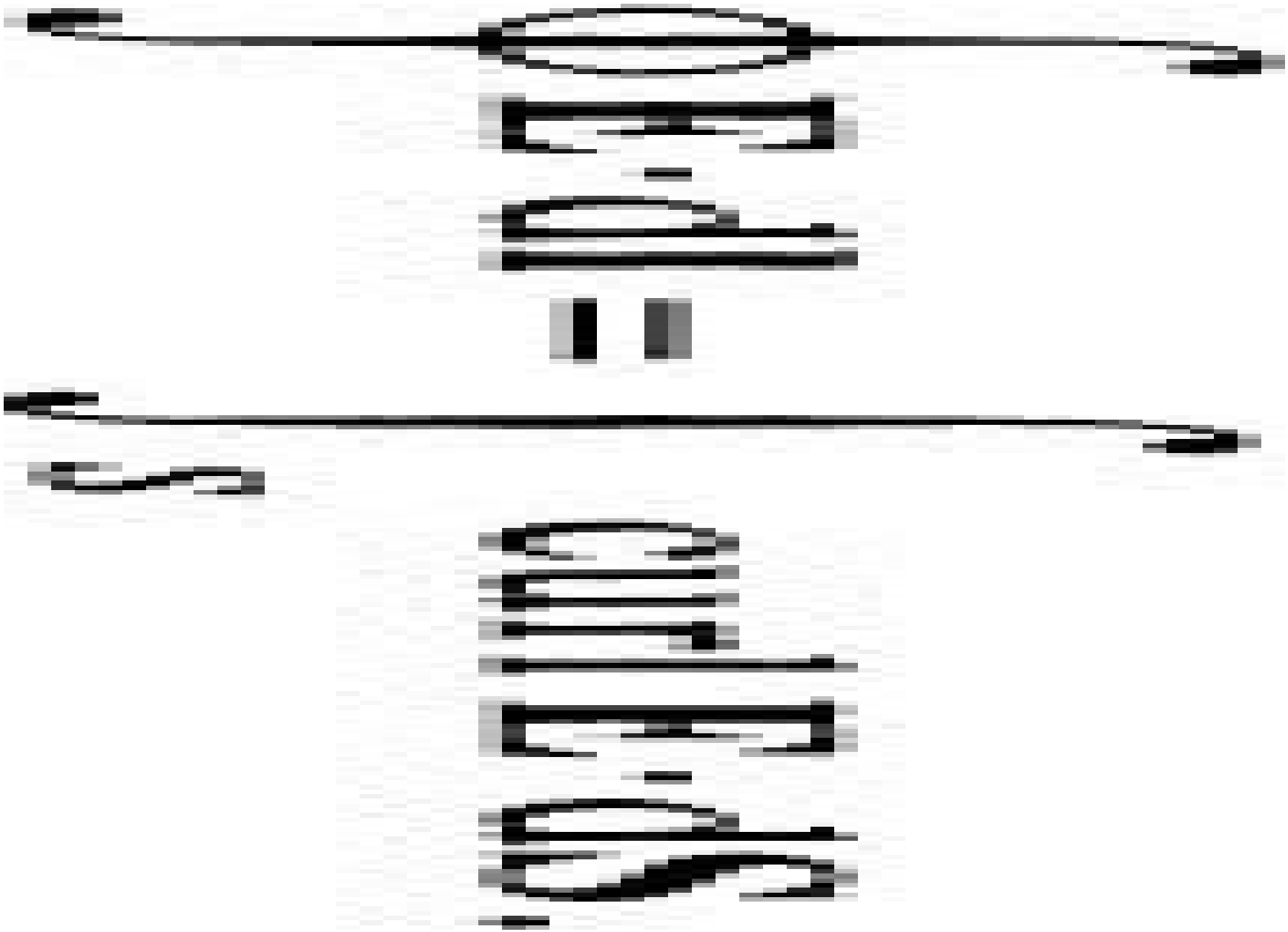
6.1.4 The differential form of Faraday's law

The integral form of Faraday's law can be reduced to a differential equation which the fields have to satisfy at every point. The procedure is very similar to that used in reducing the integral form of Ampére's law to a differential equation.

The line integral of an induced electric field around a stationary loop is related by Equation (6.13) to the rate of change with time of the magnetic flux through a surface S whose perimeter is the loop,



Using Stokes' theorem (4.45), the left-hand side of this equation can be identified with the integral over the surface S of the curl of the vector E,



We thus have

$$\int_S \operatorname{curl} \mathbf{E} \cdot d\mathbf{S} = -\frac{d}{dt} \int_S \mathbf{B} \cdot d\mathbf{S}$$

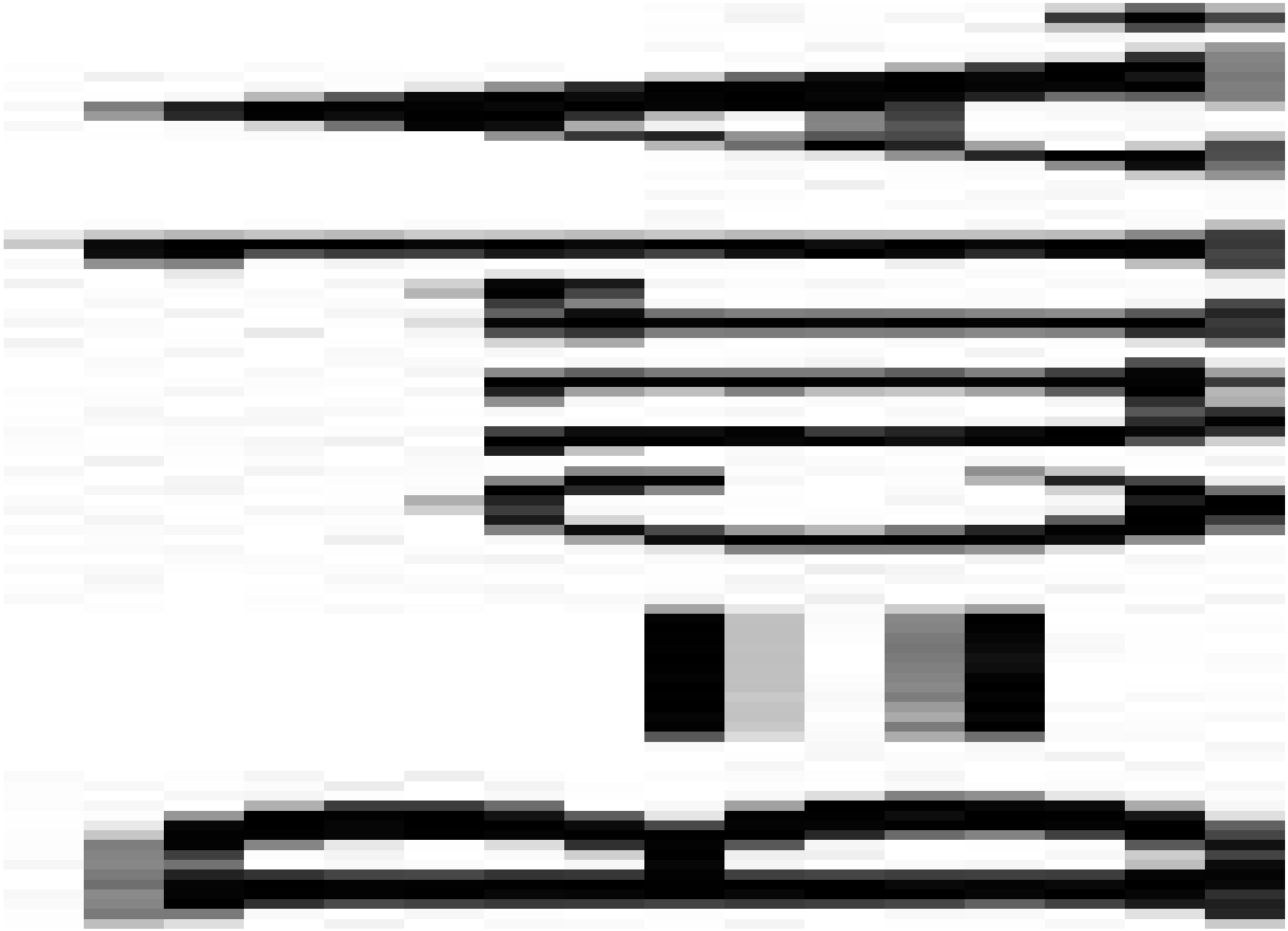
$$= \int_S -\frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{S}.$$

We note that it is permissible to change the order of integration and differentiation because we are considering a surface S which does not change shape with time and is in a fixed position. The total time differential outside the integral becomes a partial derivative under the integral sign, as we are only concerned with the changes in field B with time at the positions of the elemental areas dS. This equation is true for any surface S, and hence at all points the fields must obey the relation

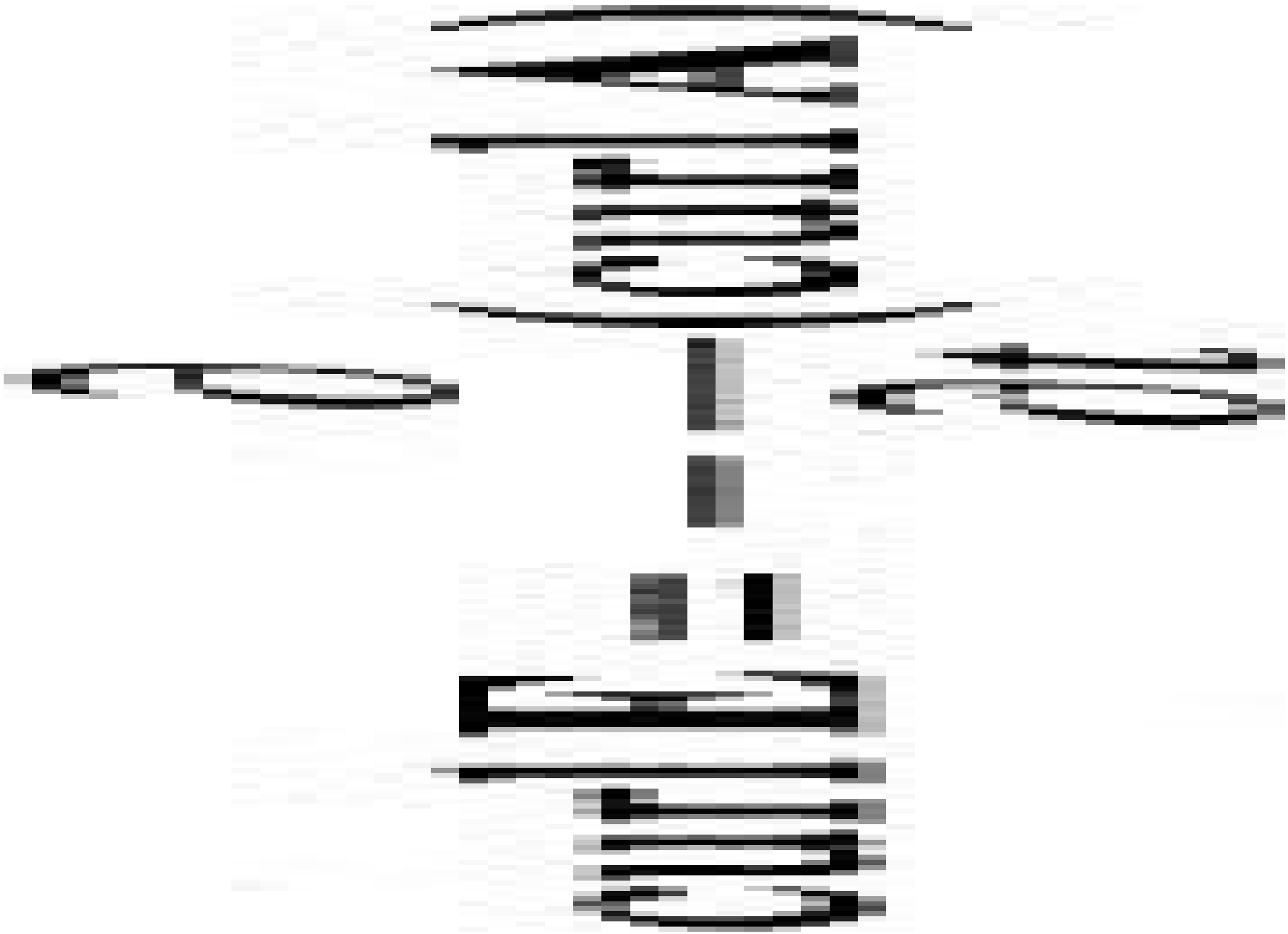
(6.17)

When magnetic fields change with time there is an induced electric field which satisfies Equation (6.17). However, the total electric field may contain a static component and the question arises as to whether the total field then continues to obey the above equation. The electrostatic field is a conservative field, as discussed in section 1.5, and its line integral around any closed loop is zero. Hence the curl of a static component of an electric field is zero everywhere, and Equation (6.17) is obeyed at all points by the total electric field. Equation (6.17) is the differential form of Faraday's law, and the field E appearing in the equation can be taken to be the total field E.

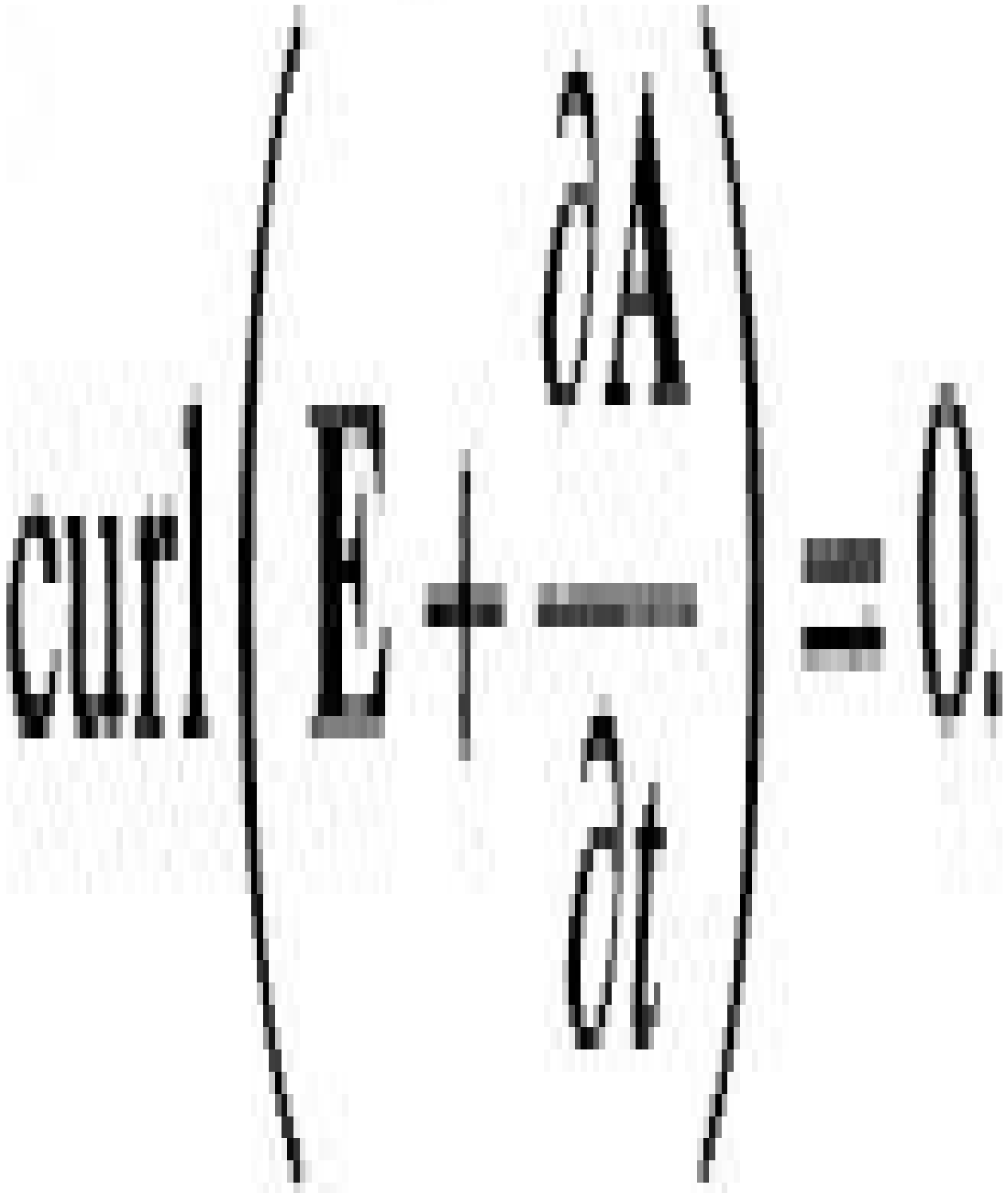
We note that since the line integral of the induced electric field around a closed loop is not zero, the total field E is not usually a conservative field and hence can not usually be derived in terms of the gradient of a scalar potential. The induced electric field, or an electric field in general, has to be derived from a more complicated potential. In section 4.5.3 we saw that since at all points the divergence of the field B was zero, we could obtain the field B by taking the curl of another vector A, the magnetic vector potential. From Equation (4.49)



Substitution of this expression for the field B into Equation (6.17) gives



or



A solution to this equation is

but the general solution is obtained by adding to the right-hand side of the above equation any function whose curl is zero. The gradient of any scalar function Φ has zero curl everywhere, and we may write

(6.18)

The function Φ reduces to the electrostatic potential only when the fields are independent of time. When the fields vary with time the potential Φ has a more general interpretation, and in such situations is usually referred to simply as the scalar potential Equation (6.18) is a general expression for the electric field in terms of the vector and scalar potentials A and ϕ. We will not make use of it now, but it is needed in some problems where charge or current distributions change rapidly with time. These problems are discussed in Chapter 13.

Let us now consider the divergence of the total electric field E. The divergence of the electrostatic field is related to the total charge density at every point by Equation (1.20). Does this equation need modification when the total field has an induced component? Experiments show that the lines of the induced electric field, like the lines of the field B, are continuous. The divergence of the induced field is therefore everywhere zero. Equations (1.20) and (2.19),

and

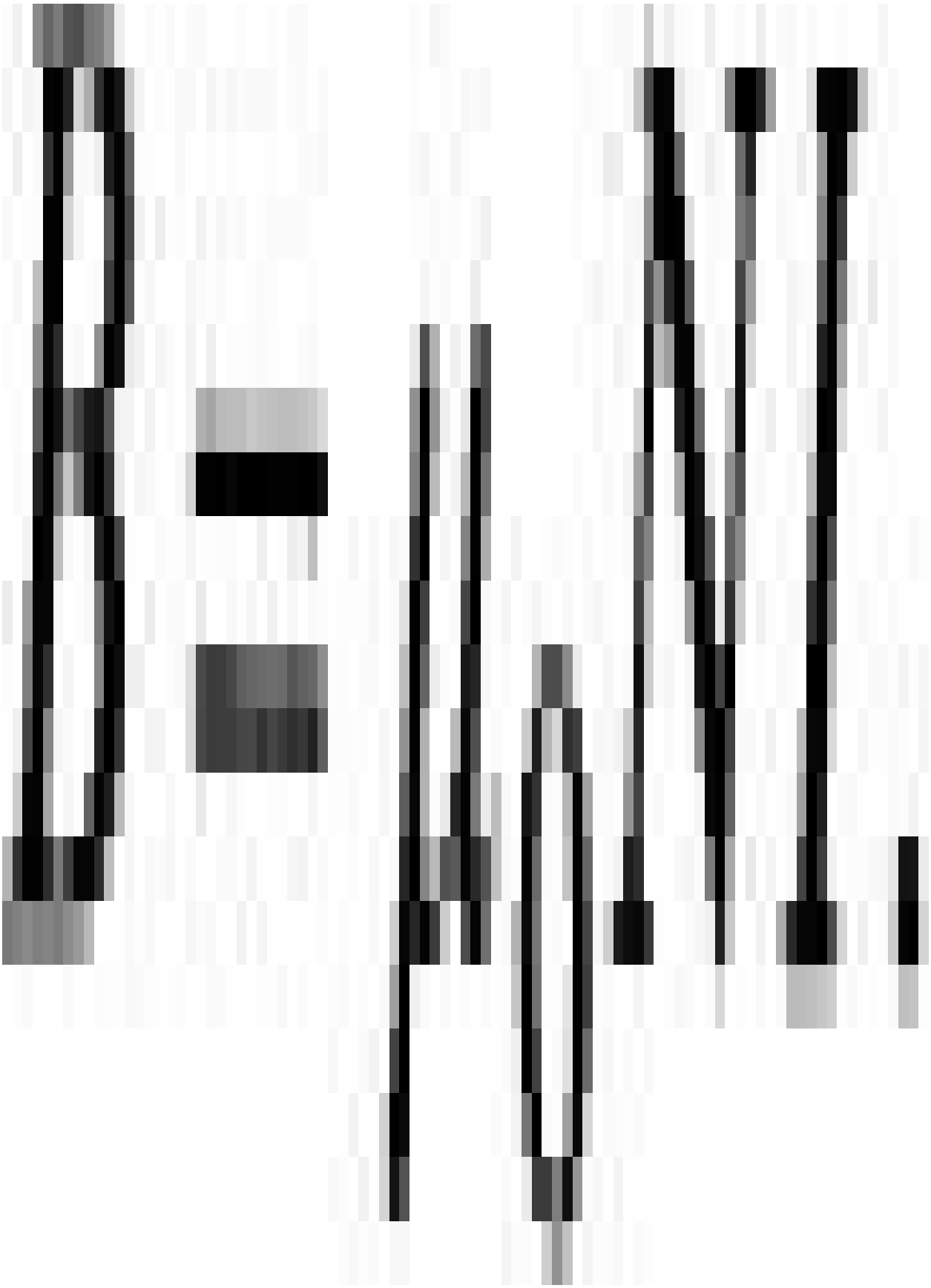
obtained in the earlier chapters on electrostatic fields thus need no modification when one is discussing the total electric field from all sources, or the total electric displacement.

6.2 SELF-INDUCTANCE AND MUTUAL INDUCTANCE

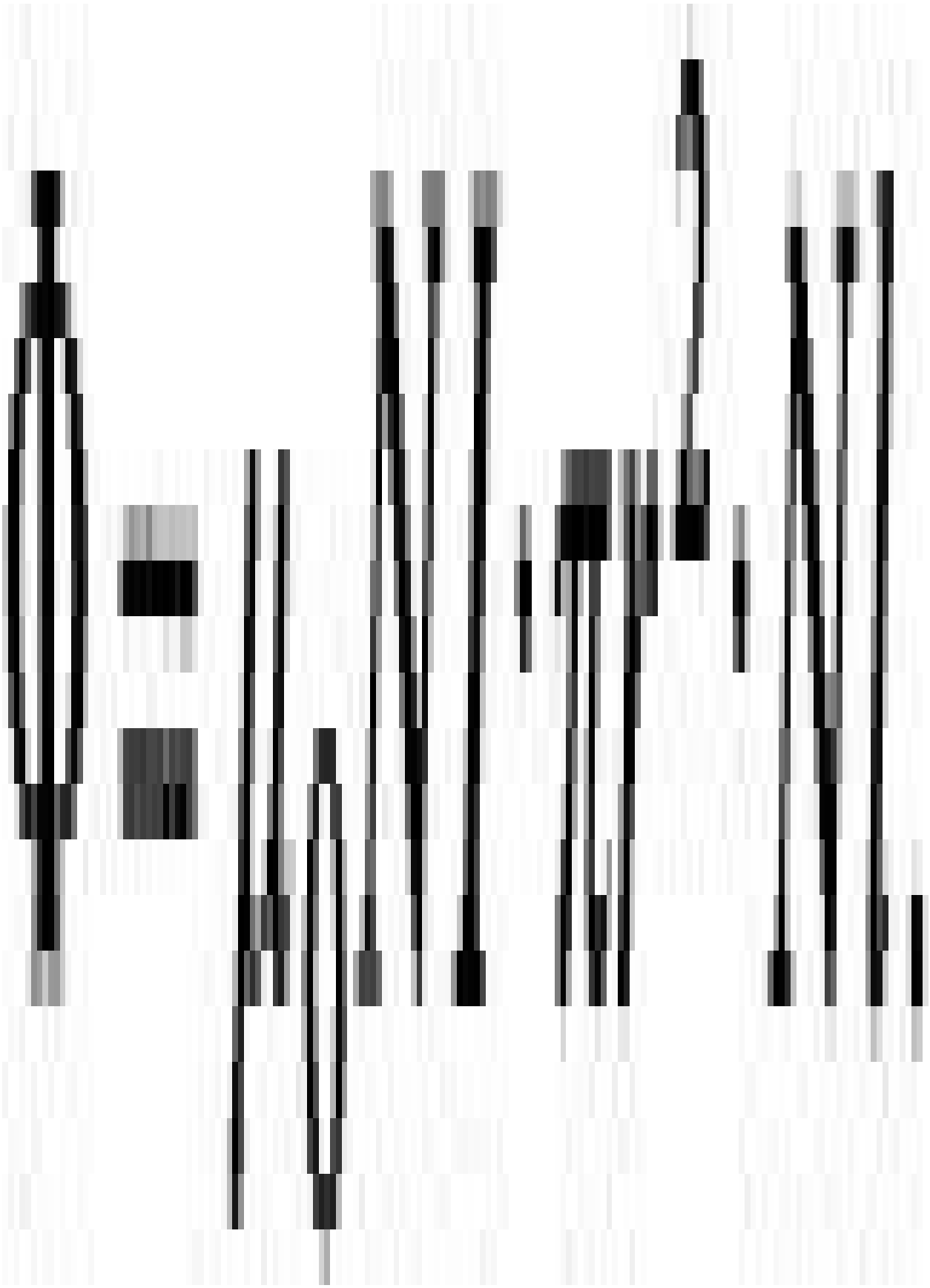
6.2.1 Self-inductance

An important application of Faraday's law is when the flux through a circuit arises solely from the current in the circuit itself. Any change in this current then induces an e.m.f. in the circuit.

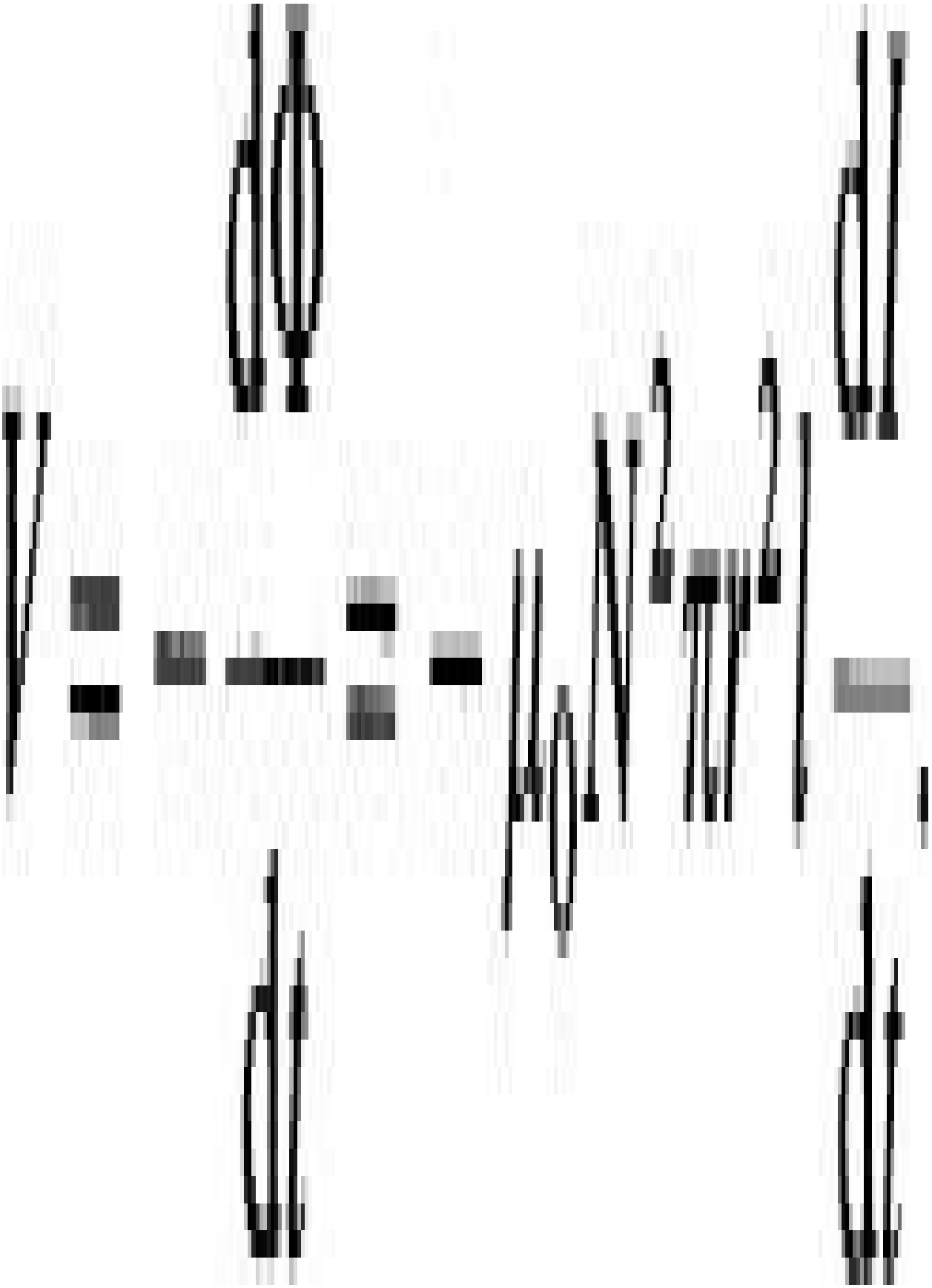
Consider a long solenoid, of length l, which has N turns per unit length and carries a current I. The magnetic field B at all points within the solenoid is given to a good approximation by Equation (4.36)



The magnetic flux through each turn of the coil is $\mu 0 NI\pi r^2$, where r is the radius of the coil. The total flux through the Nl turns on the complete solenoid is



As the current I in the solenoid changes there will be an e.m.f. V induced in the coil, which opposes the change of current:



This e.m.f. may be written as

where

(6.19)

L is called the self-inductance of the coil and depends only on the number of turns and the size of the coil.

For a coil of any description the e.m.f. V induced when the total flux Φ changes, is

$$V = -\frac{d\Phi}{dt}$$

$$= -\frac{d\Phi}{dI} \cdot \frac{dI}{dt}$$

$$= -L\frac{dI}{dt},$$

where

(6.20)

If the total flux Φ is proportional to the current I through the coil. Equation (6.20) may be written

(6.21)

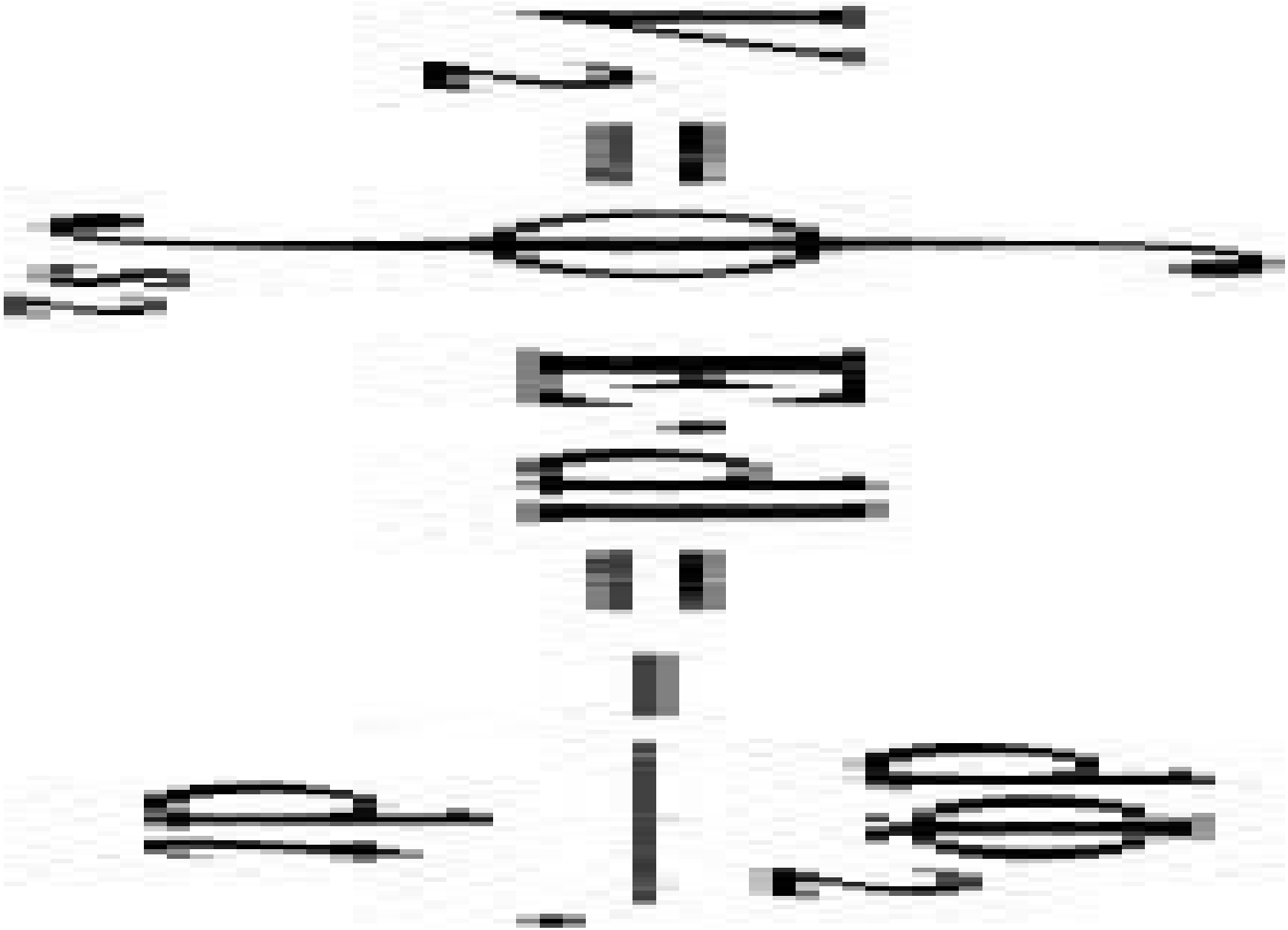
Equation (6.21) defines the self-inductance of a coil in air, where the field produced is always proportional to the current. Equation (6.20) must be used when the relation between flux and current is non-linear, as it is if the coil contains a ferromagnetic core, or if there is a contribution to the total flux from sources activated by the changing current, as there is, for example, from eddy currents in a conducting rod inside a coil. In such cases the self-inductance (and effective resistance) of the coil depend on the magnitude of the changing current I and its angular frequency ω, and Equation (6.20) defines an incremental self-inductance valid over limited ranges only.

The dimensions of inductance can be determined from Equation (6.21) and the results of section 4.6.1. They are those of mass times length squared divided by charge squared. The SI unit of inductance is called the henry, symbol H*. A self-inductance of one henry will produce e.m.f. of one volt when the current through it changes at the rate of one ampere per second. The self-inductance of a solenoid which is 50 cm long and has 5000 turns each of radius 5 cm is given by Equation (6.19) to be about 0.5 H. The henry is a very large unit, and coils used in circuits typically have inductances of the order of mH. The self-inductance of a wedding ring is about 10−8 H.

The self-inductance of a coil is important when the coil forms part of an alternating current circuit, as discussed in Chapter 7. At very high frequencies the current distribution over the cross-sectional area of the wire of the coils changes as a result of the ‘skin effect’ discussed in Chapter 11. This causes a small change of the inductance with frequency, which is normally not very important.

6.2.2 Mutual inductance

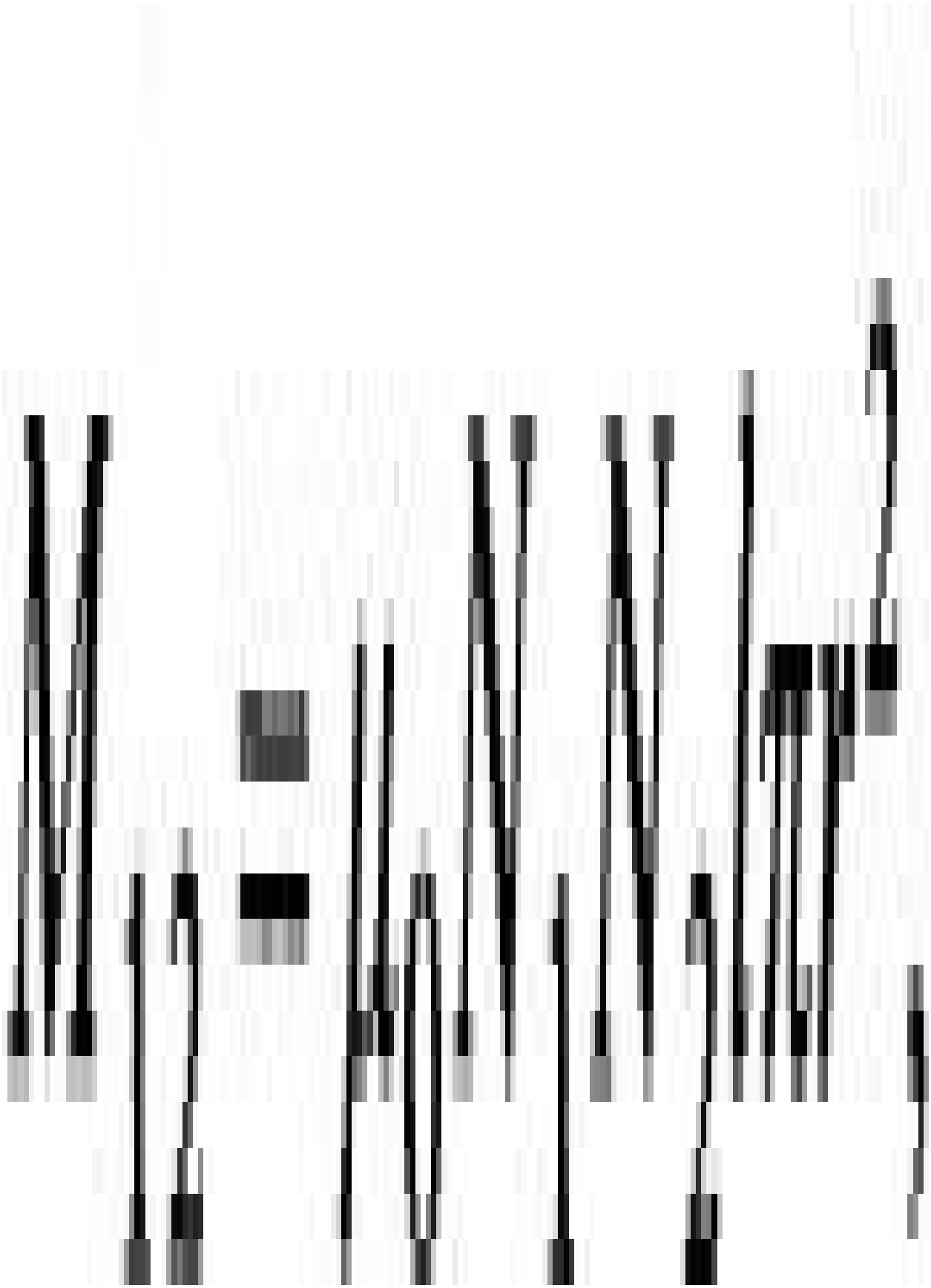
Figure 6.9 shows two long solenoids, one wound on top of the other. The length of each solenoid is l, and the common radius is r. The bottom (primary) coil has N1 turns per unit length and carries a current I1 which produces a field B1 and a total flux $\Phi 2$ through the top (secondary) coil. When I1 changes $\Phi 2$ changes and an e.m.f. V2 is induced in the secondary given by



In this formula the secondary loop s2 has to be traversed in the direction which makes $\Phi 2 = \int_{s2} B1.dS$ positive, in a manner consistent with the discussion of Lenz's law in section 6.1.2. We then have

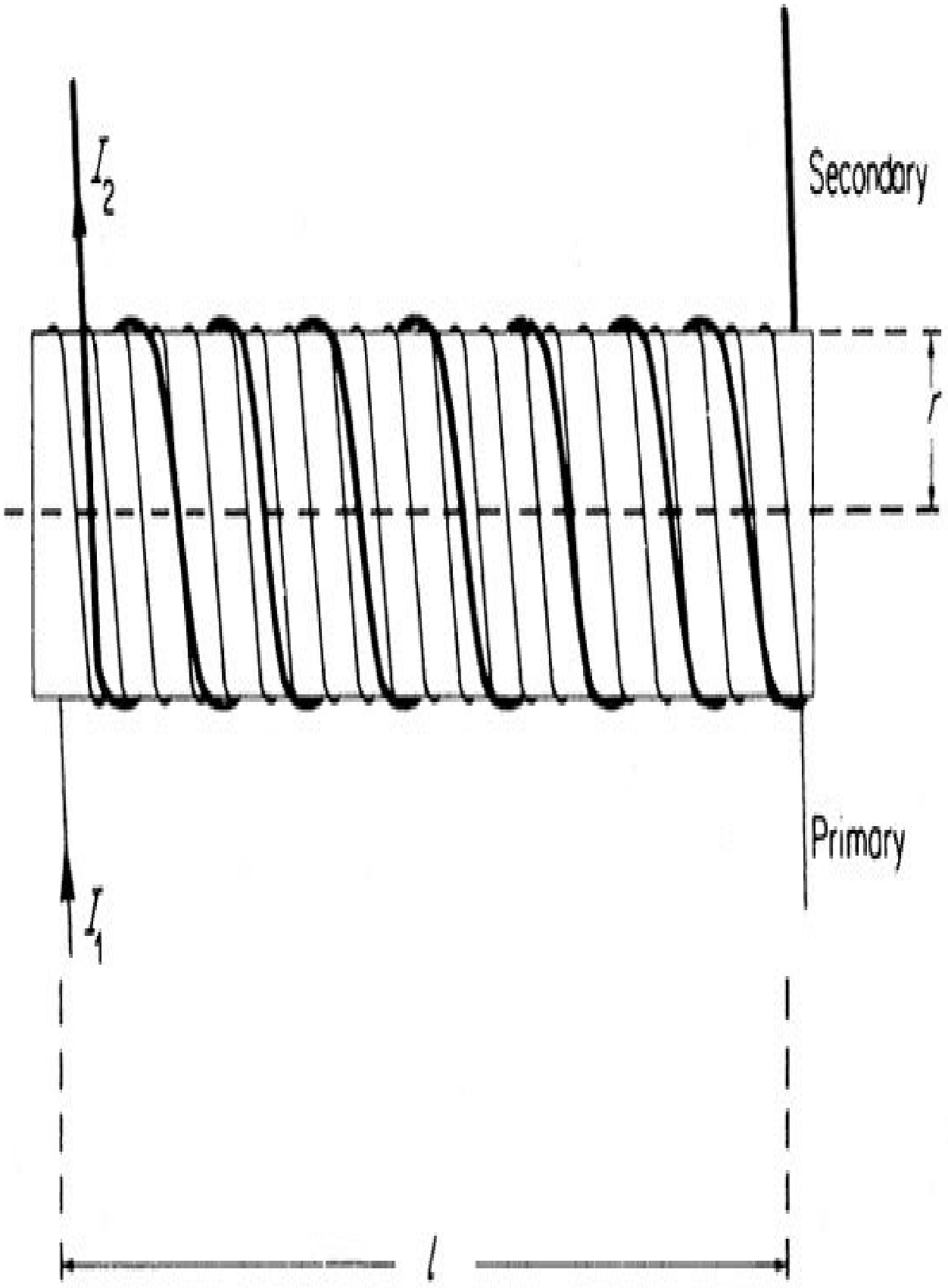
(6.22)

with



where N2 is the number of turns per unit length on the top coil. For I1 increasing so that dI1/dt is positive, V2 is negative and the induced current which is trying to maintain Φ2 constant is in the direction shown on Figure 6.9.

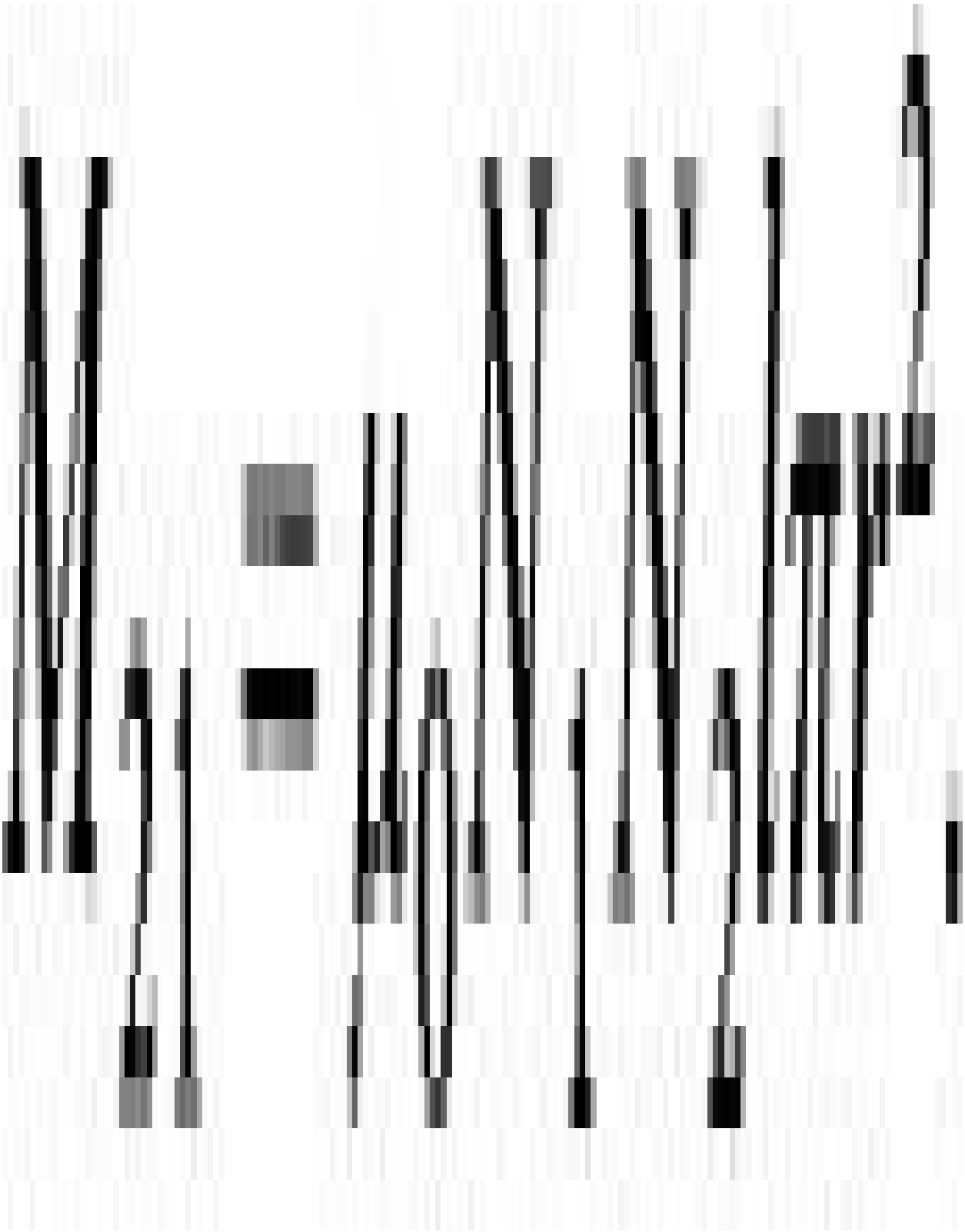
Figure 6.9. A mutual inductance made by winding one long solenoid on top of another.


I_2
Secondary
r
Primary
I_1
l

If we reverse the roles of the top and bottom coils by letting a changing current I2 in the top coil induce an e.m.f. V1 in the bottom coil it can be seen that

(6.23)

where



The quantities M12 and M21 thus have the same value and can be given the symbol M. M is called the mutual inductance of the two coils, and depends only on their construction and geometry.

The result that the induced e.m.f.s in the two coils are given by Equations (6.22) and (6.23) with M12 = M21 = M is always true. The mutual inductance between two coils is equal to the flux through one due to unit current in the other if the coils are in air, where the field produced is proportional to current.

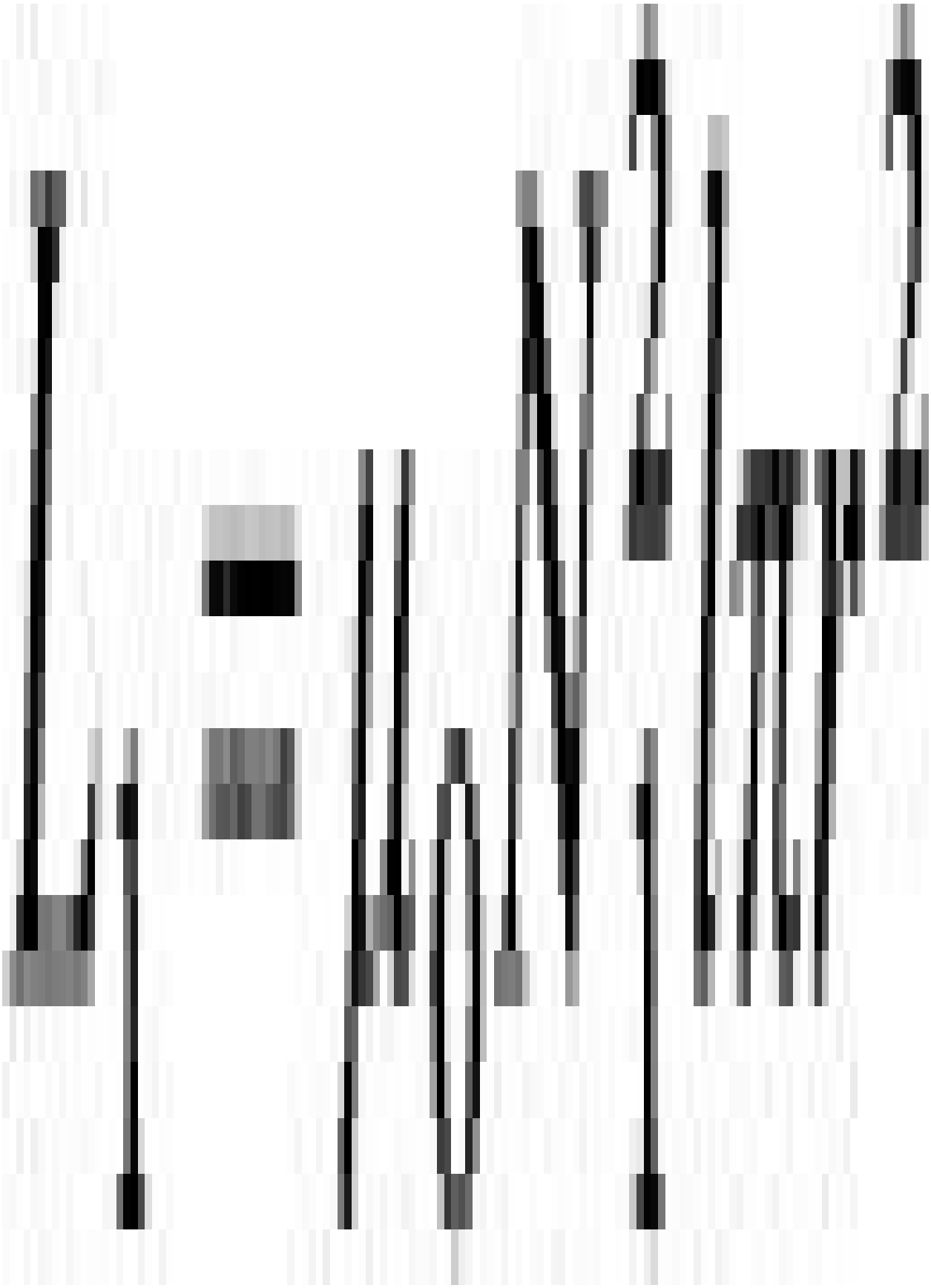
If the flux Φ2 through the secondary coil of a mutual inductance is not proportional to the current I1 in the primary coil, the voltage V induced in the secondary when the current I1 changes can be written

$$V = -\frac{d\Phi_2}{dI_1}\cdot\frac{dI_1}{dt}$$

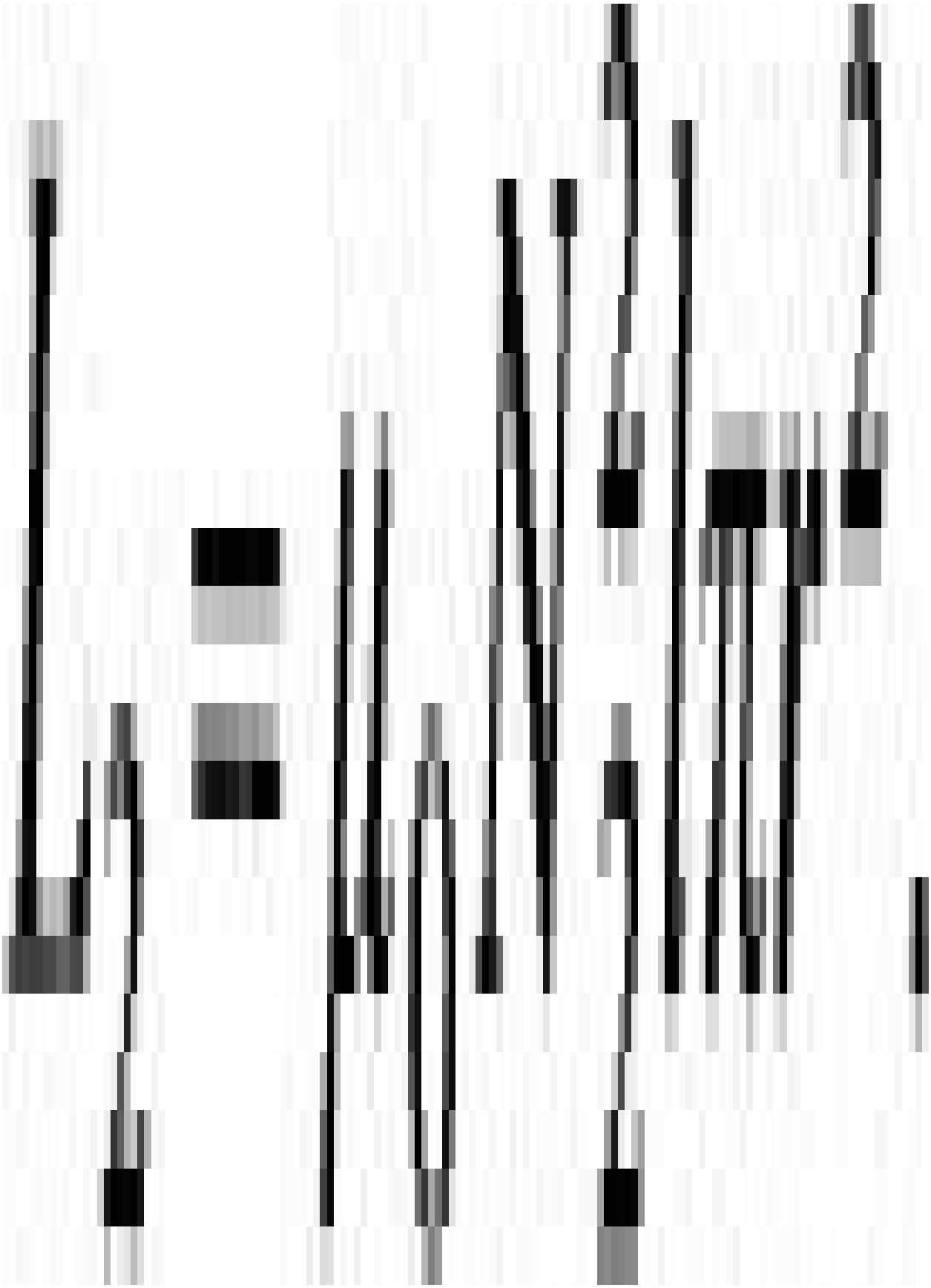
$$= -M\frac{dI_1}{dt}.$$

The mutual inductance $M = d\Phi 2/dI1$ varies, and depends on the non-linear relationship between flux and current. The incremental mutual inductance used for coils wouncl on iron formers depends on the B–H curve for the iron and on the effects of currents induced in the iron when the current I1 changes. Pairs of coils wound on iron are usually called transformers; the properties of transformers are discussed in Chapter 8.

The self-inductance of the bottom coil of Figure 6.9 is given by



from Equation (6.19). The self-inductance of the top coil is



Hence the mutual inductance of the two coils can be written

(6.24)

This result depends on the assumption that all of the flux produced by one coil passes through the other coil. In a real situation this is not so, and the mutual inductance is less than that given by Equation (6.24). We may write

(6.25)

where the dimensionless number k, called the coefficient of coupling, is less than or equal to unity.

A car ignition coil is a familiar example of a mutual inductance. In a typical example two insulated coils, one of 16 000 turns the other of400 turns, are wound over each other. The length of each coil is 10 cm and the turns all have roughly the same radius of 3 cm. A current of 3 A is passed through the primary coil and broken in about 10^{-4} seconds, so that $dI/dt \sim 3 \times 10^{4}$A s^{-1}. The voltage induced in the secondary circuit is given by Equation (6.22), and is about 6000 V. This causes a spark to jump across the gap in a spark plug and ignite a gasoline–air mixture.

The mutual inductance of two coils can be calculated with the aid of the Biot–Savart law (4.32) by assuming that the wire of the coils is vanishingly thin, the condition for validity of Equation (4.32). The expression for the mutual inductance is given in the answer to Problem (4.16). It is interesting to note that

the same technique cannot be directly applied to the calculation of self-inductance. If a current flows in a vanishingly thin wire coil (the situation to which the Biot–Savart law applies) the flux through the coil is infinite. In order to calculate the self-inductance of a real coil an integration must be performed over the current filaments constituting the current in the wire. Each filament is vanishingly thin but carries a vanishingly small current, and if the integration is done carefully no singularities arise. For accurate calculations of mutual inductances the same procedure has to be adopted.

6.3 ENERGY AND FORCES IN MAGNETIC FIELDS

In this section we consider magnetic energy. In electrostatics the work done in bringing a system of charges to their final configuration is stored as energy in the electrostatic field. In magnetostatics some of the work done in setting up a current distribution is stored as energy in the magnetic field. We will derive expressions for the total magnetic energy associated with a system of currents and apply the results to some practical situations.

6.3.1 The magnetic energy stored in an inductor

Let us suppose that at some time t = 0 an inductive coil is placed across the terminals of a battery of e.m.f. V. We will work out the current IT in the coil at a time T. The coil is equivalent to a self-inductance L which has negligible resistance, in series with a resistor of resistance R. At any time the voltage drop across the self-inductance is equal to L dI/dt where I is the current, and the voltage drop across the resistor is RI, hence

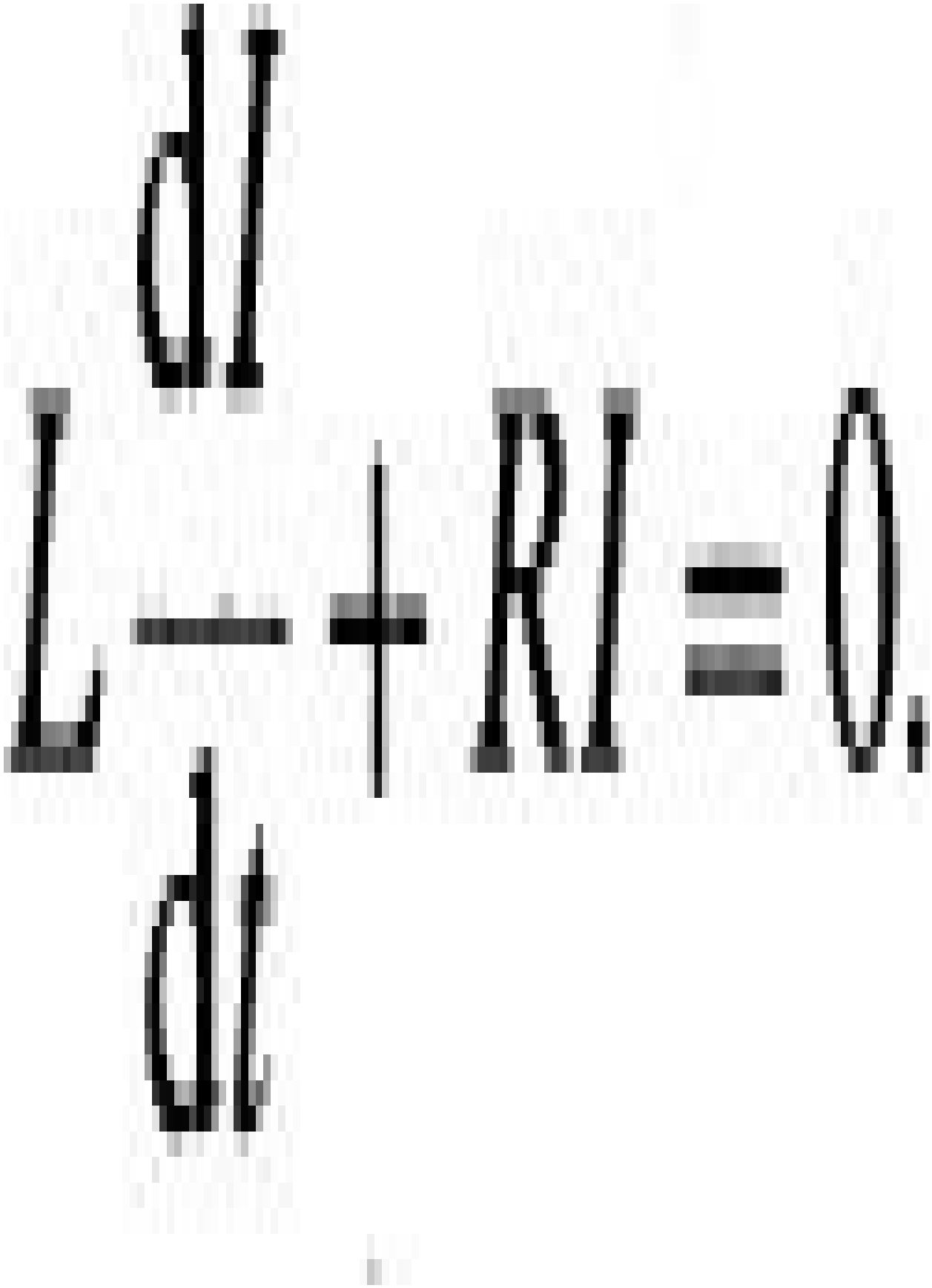
The total work WB done by the battery in raising the current to the value IT is given by

$$W_B = \int_0^T V I \, dt$$

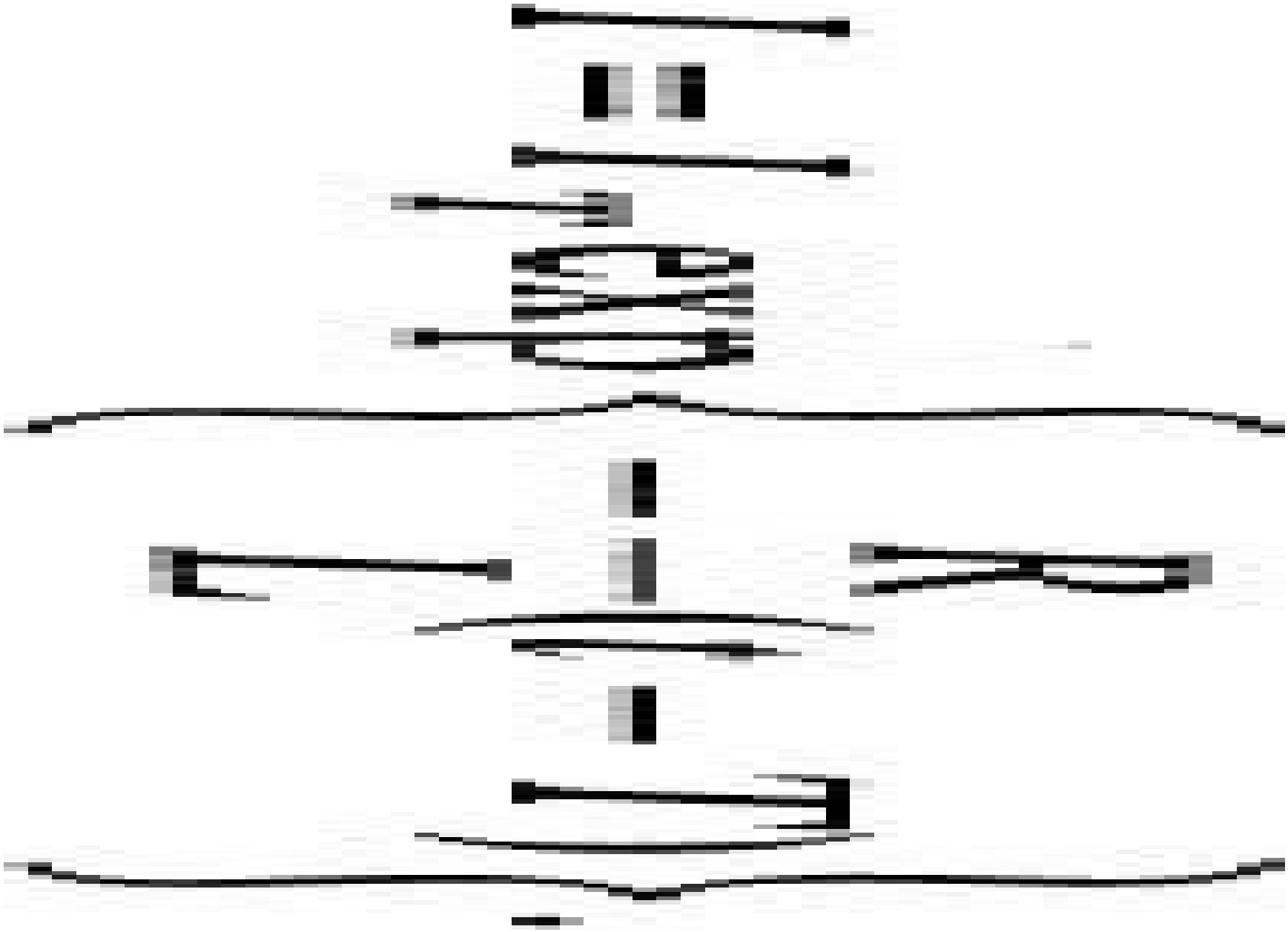
$$= L \int_0^T I \frac{dI}{dt} \, dt + R \int_0^T I^2 \, dt$$

$$= \tfrac{1}{2} L I_T^2 + R \int_0^T I^2 \, dt.$$

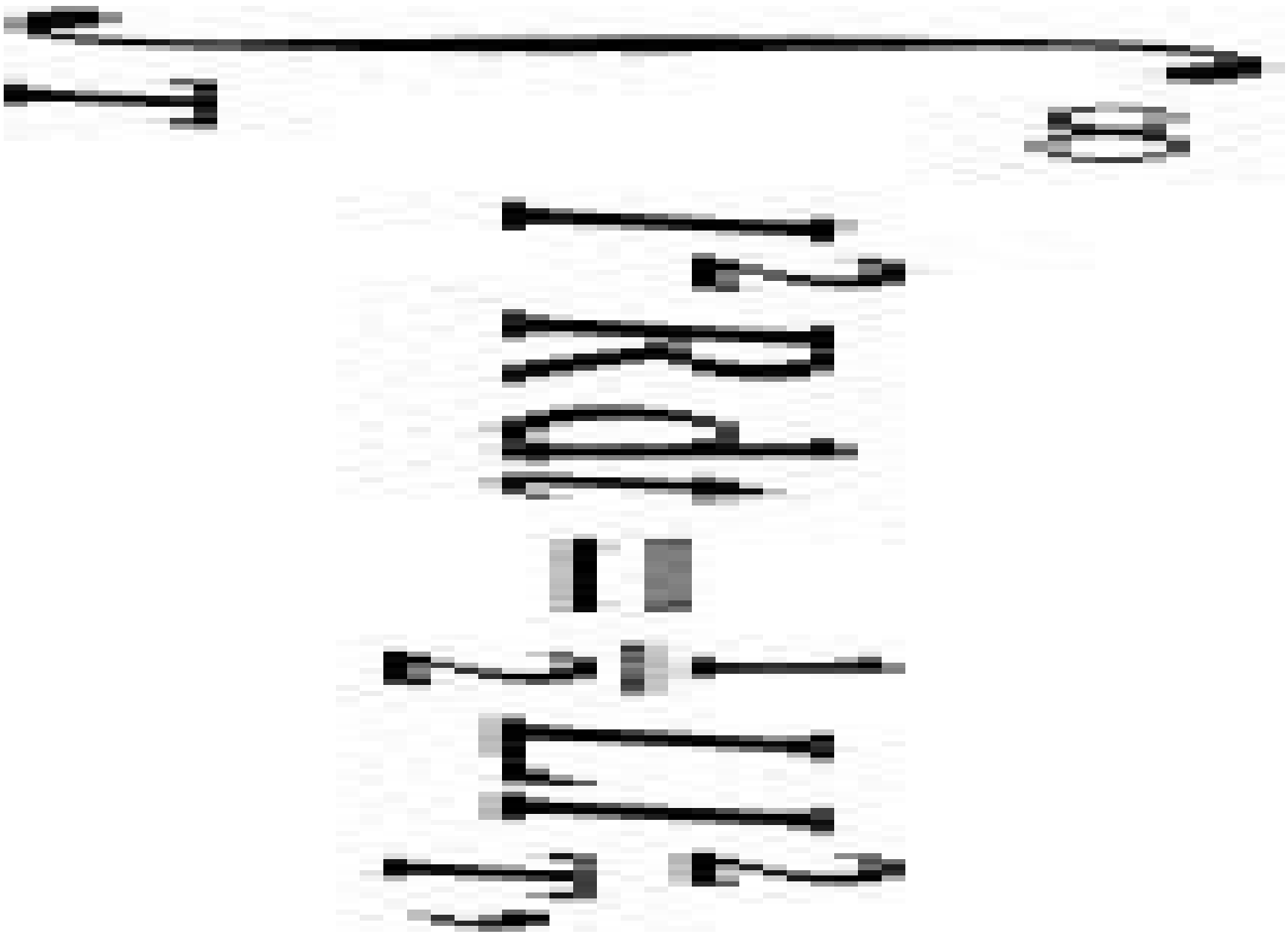
The second term on the right-hand side of this equation represents the irreversible conversion of electrical energy into heat. The first term is an amount of energy stored in the inductance after time T, which can be recovered when the magnetic field is switched off. If at time T the battery is removed and replaced by a short circuit, at all times later than T the following equation is obeyed:



Solving this equation for the current I with the initial condition that at time t = T, the current I = IT, we have



The current decays away exponentially and dissipates energy as heat in the resistor. The total energy appearing as heat after the removal of the battery is given by the integral

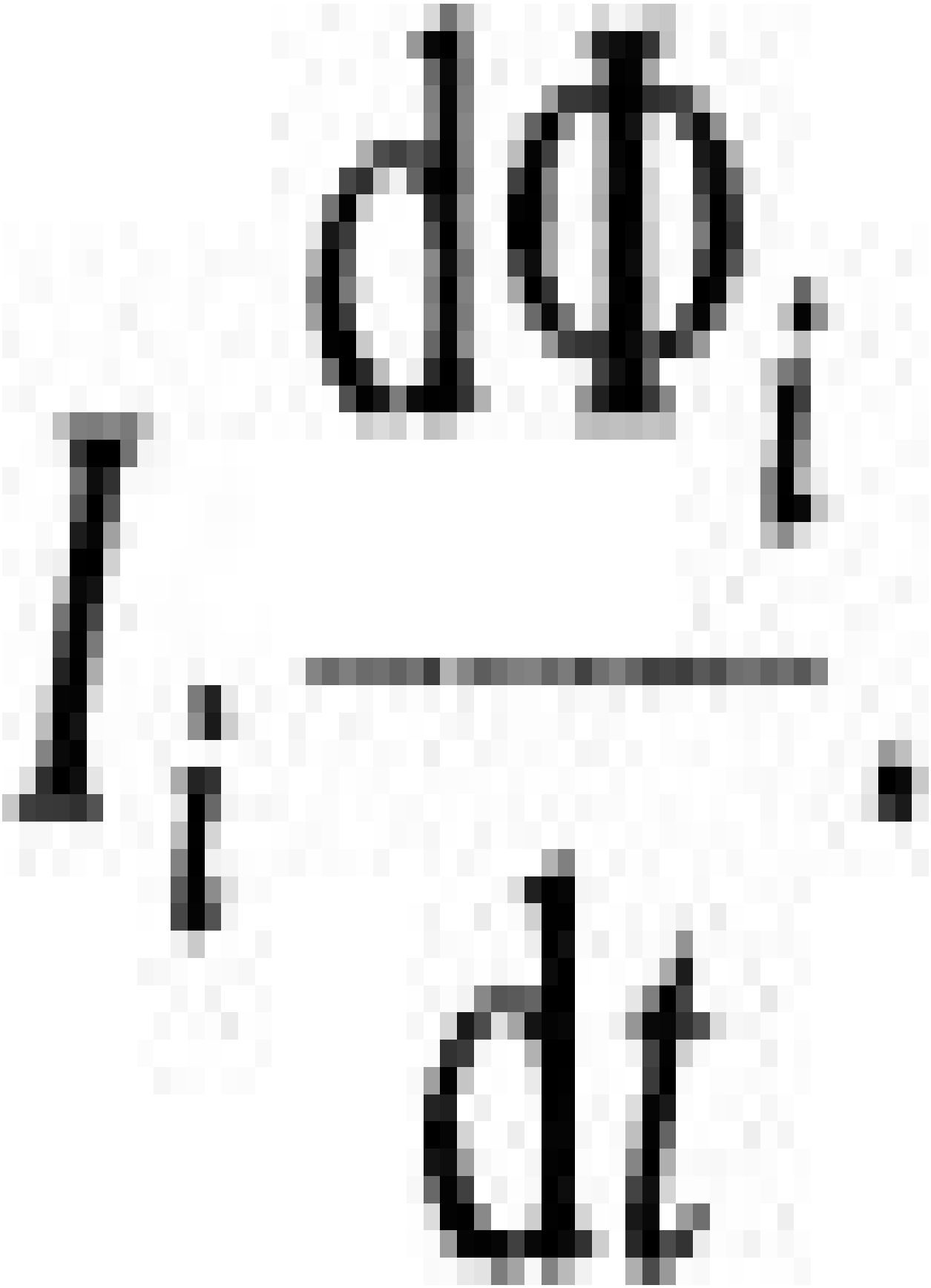


and the heat energy appearing is equal to the energy stored in the inductance. The energy may be considered to be stored reversibly in the magnetic field caused by the current. An inductance is thus a device for storing magnetic energy, analogously to the manner in which a capacitance stores electrostatic energy.

6.3.2 The total magnetic energy of a system of currents

In the same way as the total energy stored in an electrostatic field can be expressed in terms of the electric fields E and D, the total energy stored in a magnetic field can be related to the magnetic fields B and H. In order to determine this relation a necessary step is to calculate the total magnetic energy stored in a system of circuits through which steady currents pass.

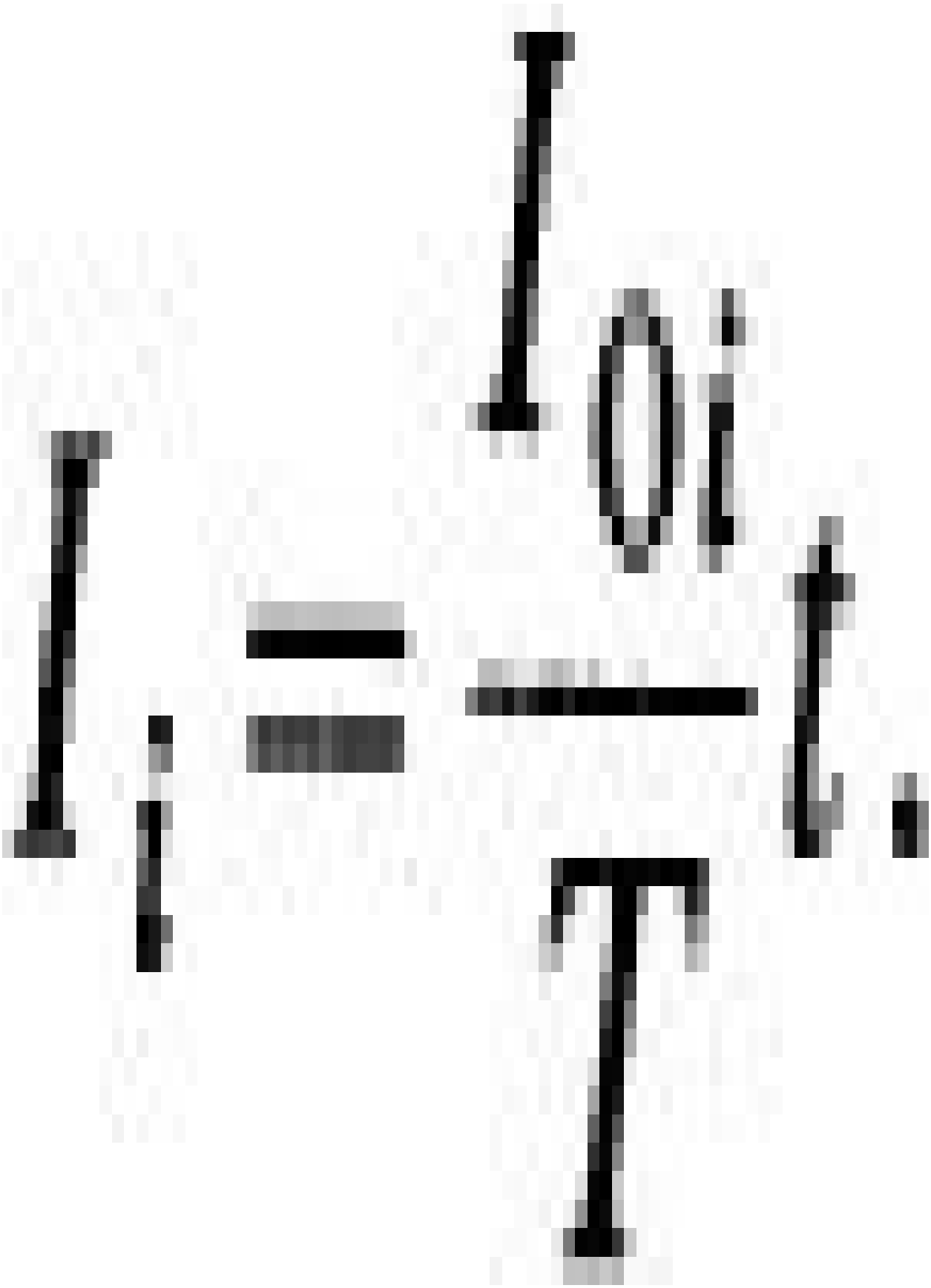
Suppose there are n stationary circuits and that at an instant of time t the current through the ith circuit is Ii and the magnetic flux through the circuit is Φi. The instantaneous induced e.m.f. in the ith circuit is equal to $-d\Phi i/dt$ and the rate at which work is being done against this induced e.m.f. by the battery supplying the current Ii is given by



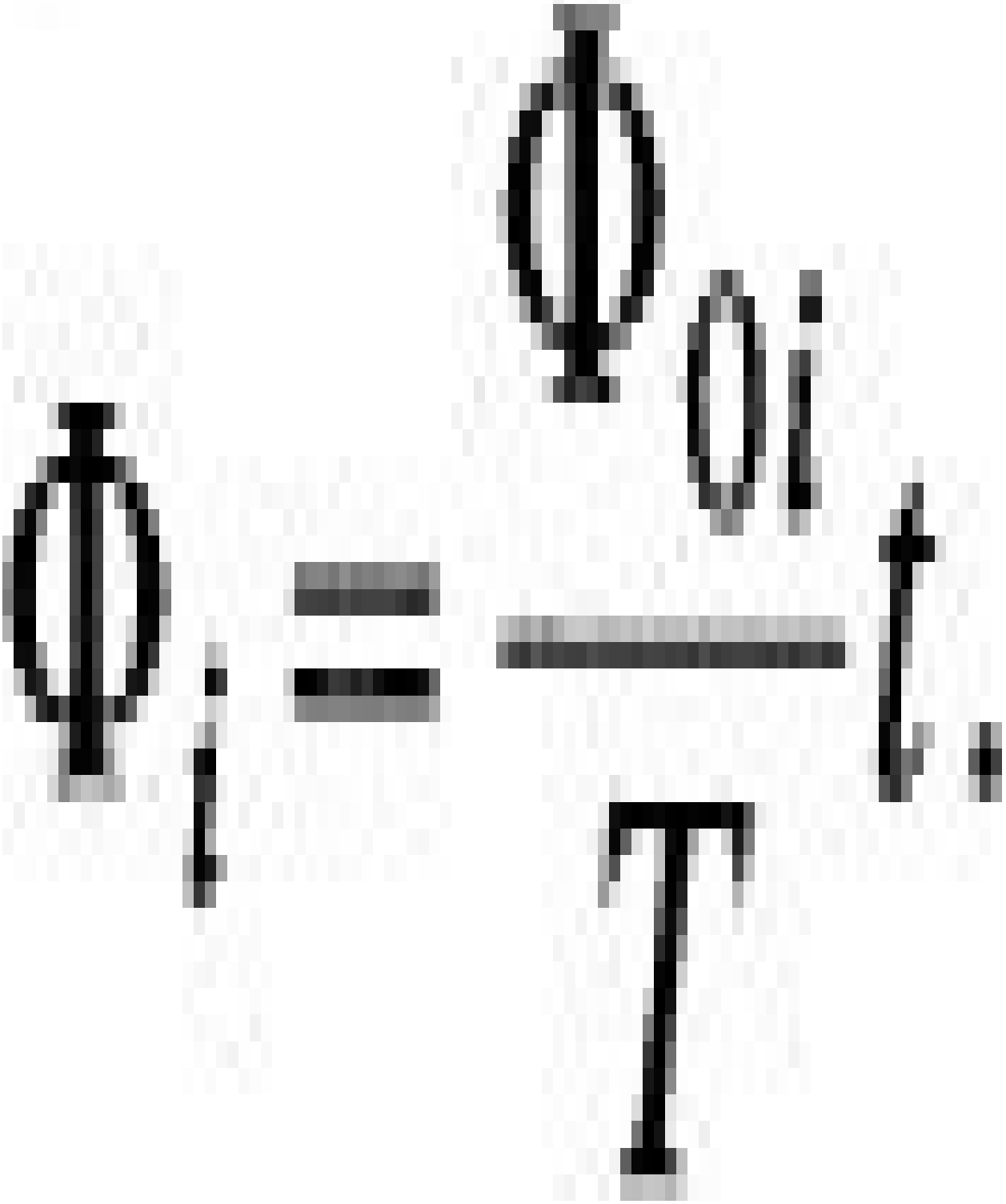
The total work d Wb done by all the batteries against the induced e.m.f.s in time dt is given by

(6.26)

If the circuits remain fixed in position during the flux changes the value of the work d Wb given by this equation is equal to the increase d U in the total magnetic energy stored in the system*. To determine the total magnetic energy U stored in a system of currents I0i flowing in circuits which have magnetic fluxes Φ0i through them, Equation (6.26) has to be integrated over the time T it takes to establish the currents. The energy U is independent of the way in which the currents are built up to their final values I0i, and so, for the purposes of calculating the energy, let us assume that they are all built up together linearly with time. In this case the current Ii at time t is given by



If the circuits are situated in linear materials the fluxes Φi are proportional to the currents and



The restriction on the validity of this equation is important; the equation would not apply for example if a ferromagnetic material were in the space between two of the circuits. The total energy U is now determined by integrating Equation (6.26) over the time T, and using the above equations for Ii and Φi,

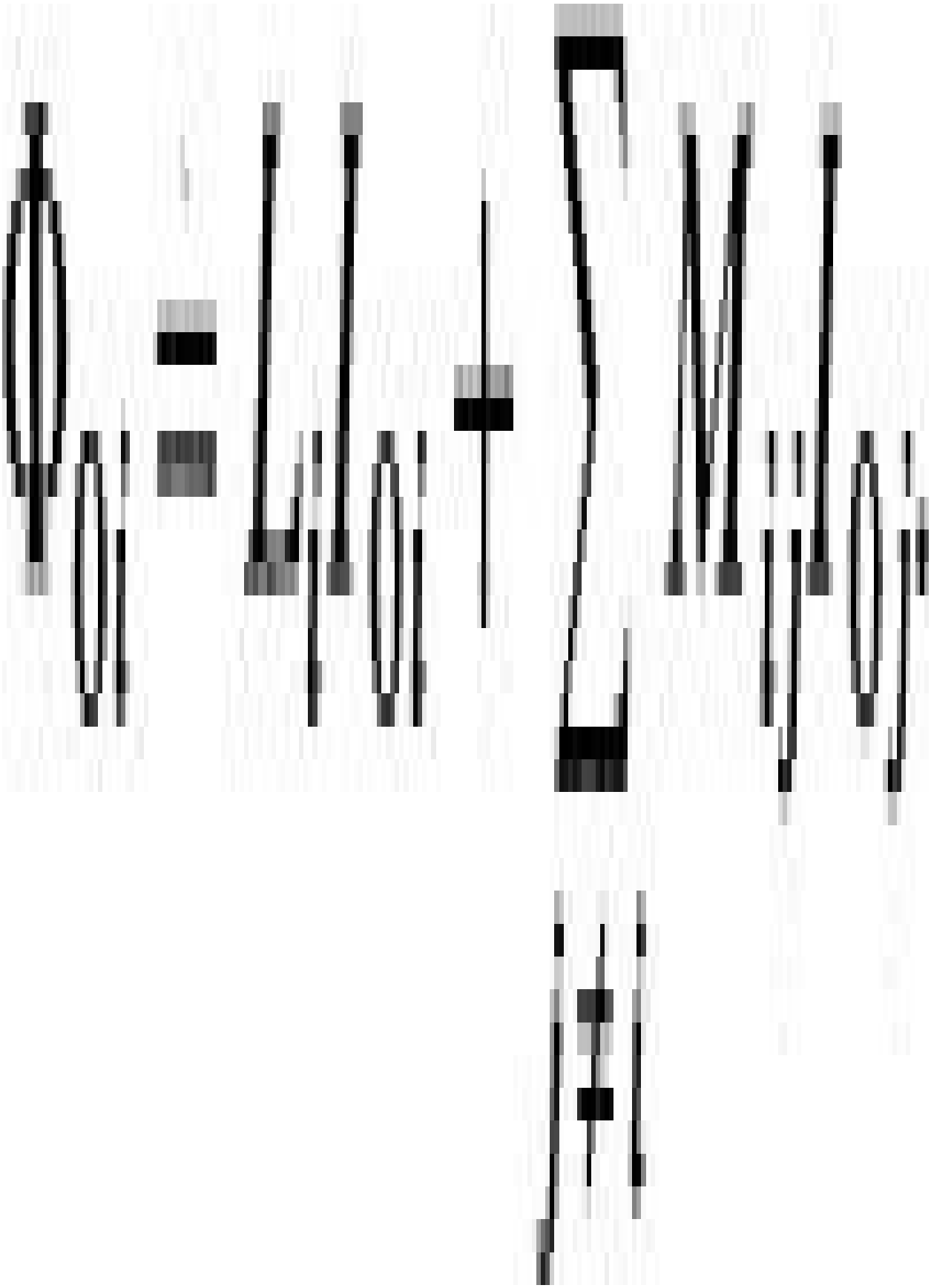
$$U = \sum_{i}^{n} \int_{0}^{T} I_i \frac{d\Phi_i}{dt} dt$$

$$= \sum_{i}^{n} \frac{I_{0i}\Phi_{0i}}{T^2} \int_{0}^{T} t\, dt.$$

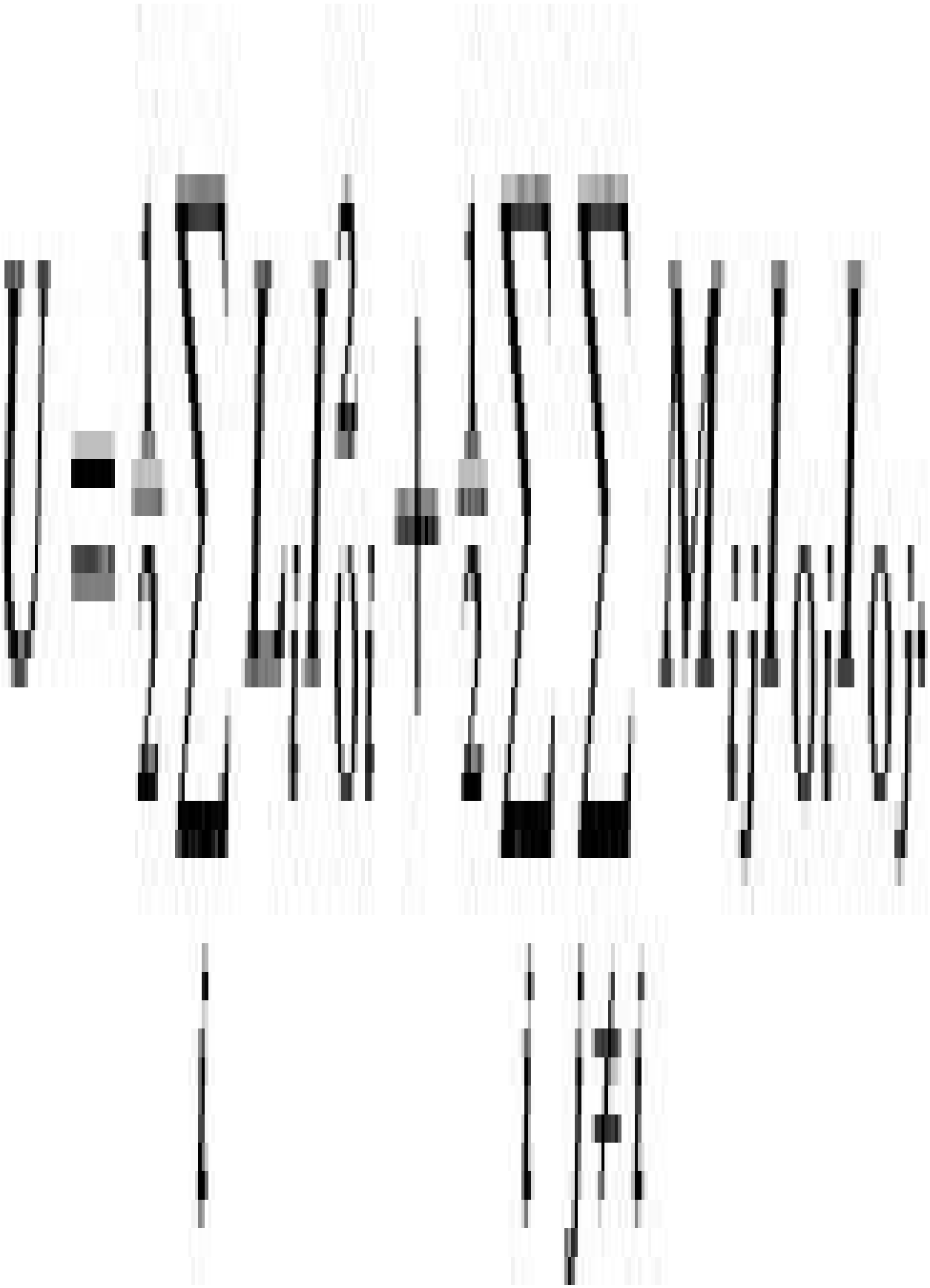
On integration this becomes

(6.27)

The quantities Φ0i can be written in terms of the self-inductances Li of the circuits and the mutual inductances Mij between the circuits:



On substitution of this expression into Equation (6.27) we obtain



For just two circuits

(6.28)

since M12 = M21. The sign of the middle term may be either positive or negative depending on whether the current in one coil increases or decreases the flux through the other coil, i.e. on whether the fields from the two coils add up or not.

6.3.3 The potential energy of a coil in a field and the force on the coil

Let us now consider the torques and forces on a circuit which carries a fixed current I1 and is placed in an external field arising from a fixed current I2 in another coil. The same problem was discussed in section 4.6.1, where expression (4.52) was obtained for the potential energy UP of coil 1 in the field B2 of coil 2,

(6.29)

where S1 is a surface whose rim is the contour of coil 1. In terms of the mutual inductance M12 between the two coils, the magnetic flux through coil 1 is given by

(6.30)

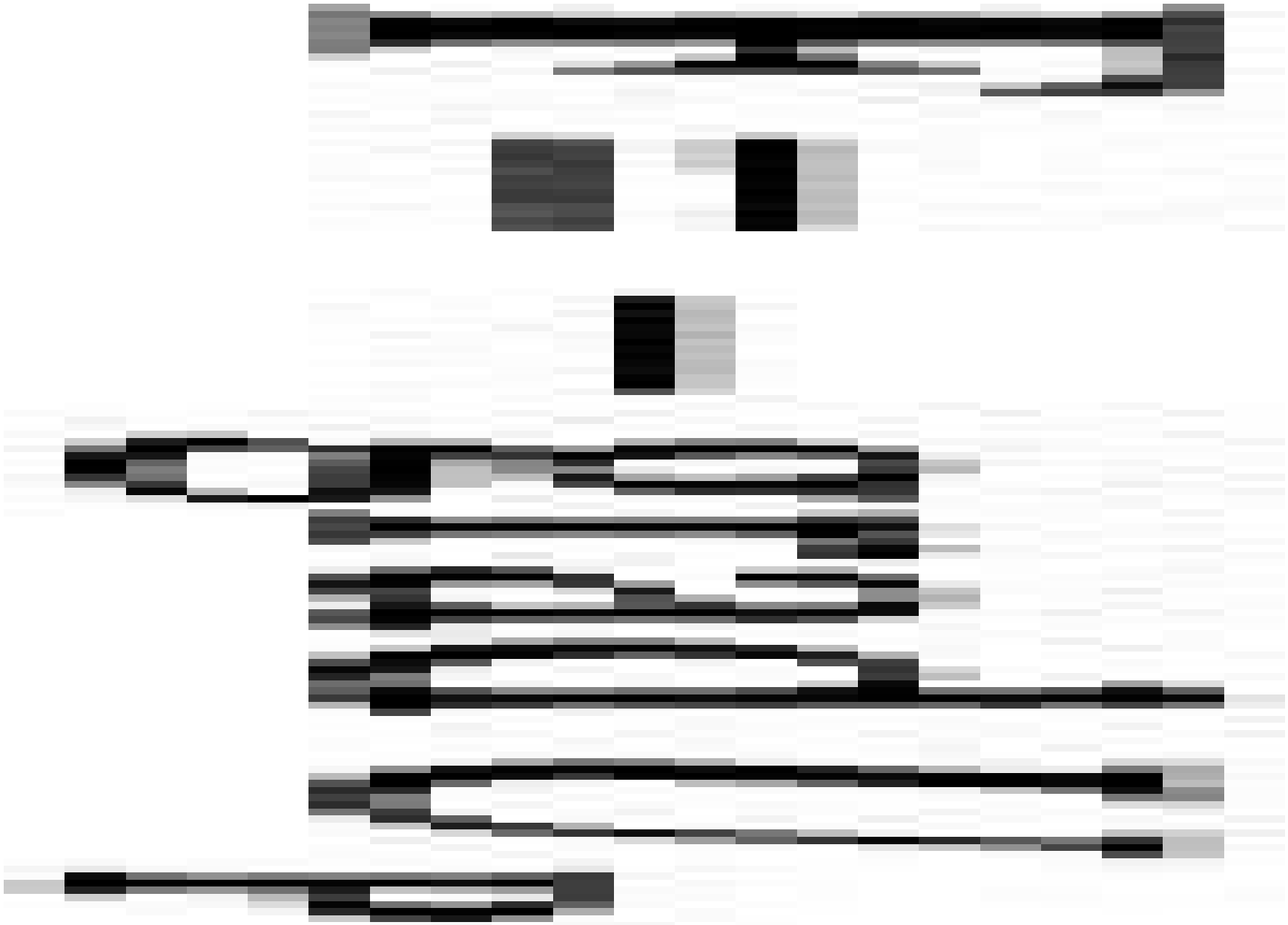
The sign in Equation (6.30) is the same as in Equation (6.28) and is determined by requiring the direction of the vectors dS to be given by the direction of the current I1 and the right-hand screw rule. Combining Equations (6.29) and (6.30),

(6.31)

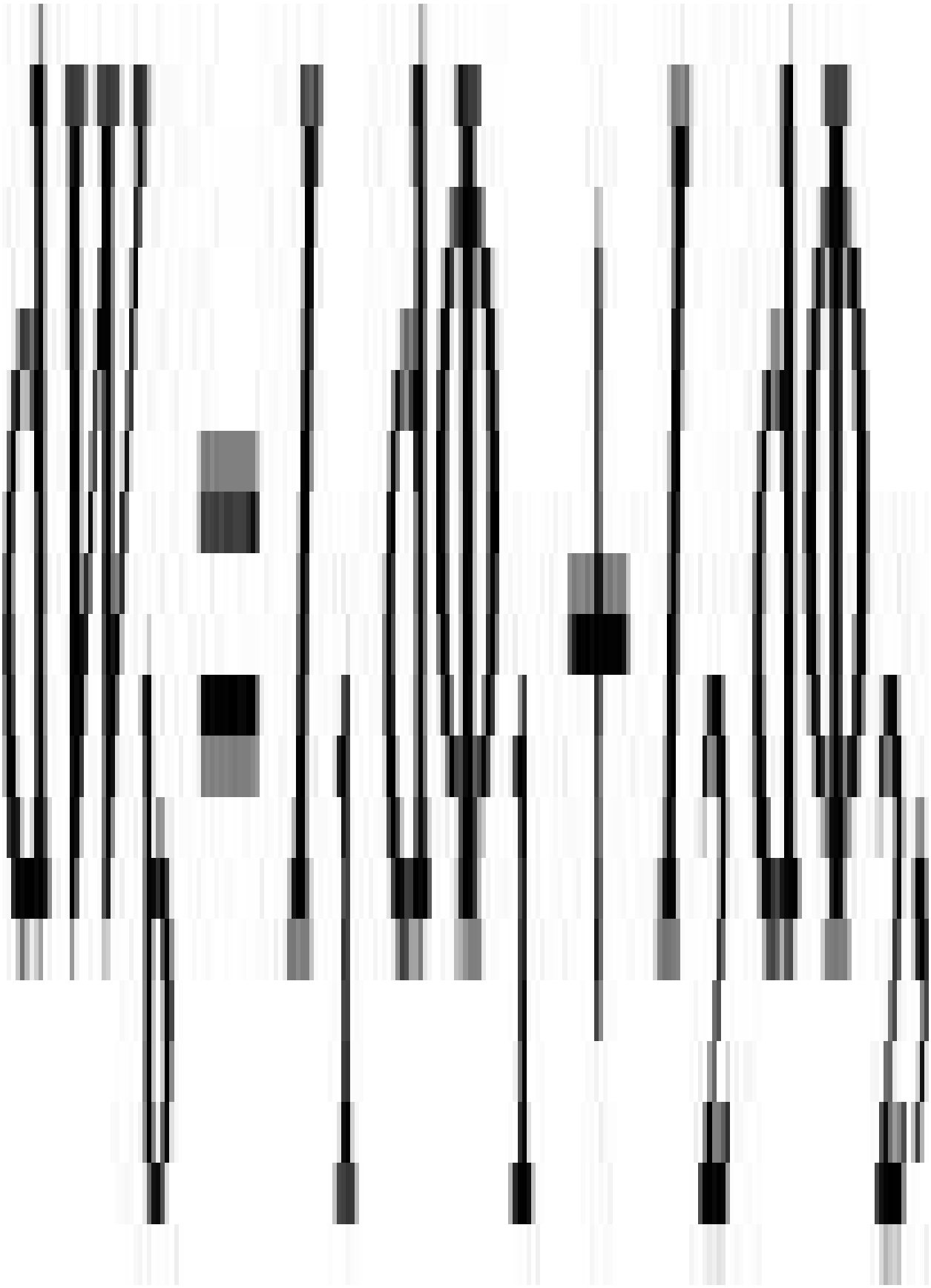
The total magnetic energy of the system, given by Equation (6.28) contains terms representing self-energies and involving self-inductances, and an interaction energy ±M12I1I2. The potential energy of one coil in the field of the other, given by Equation (6.31), is equal to M12I1I2. The potential energy UP is related to the mechanical work done by the system if a coil moves under the influence of the magnetic forces. If one coil undergoes a displacement dr under the influence of a magnetic force F, the mechanical work done by the system dW is given by

(6.32)

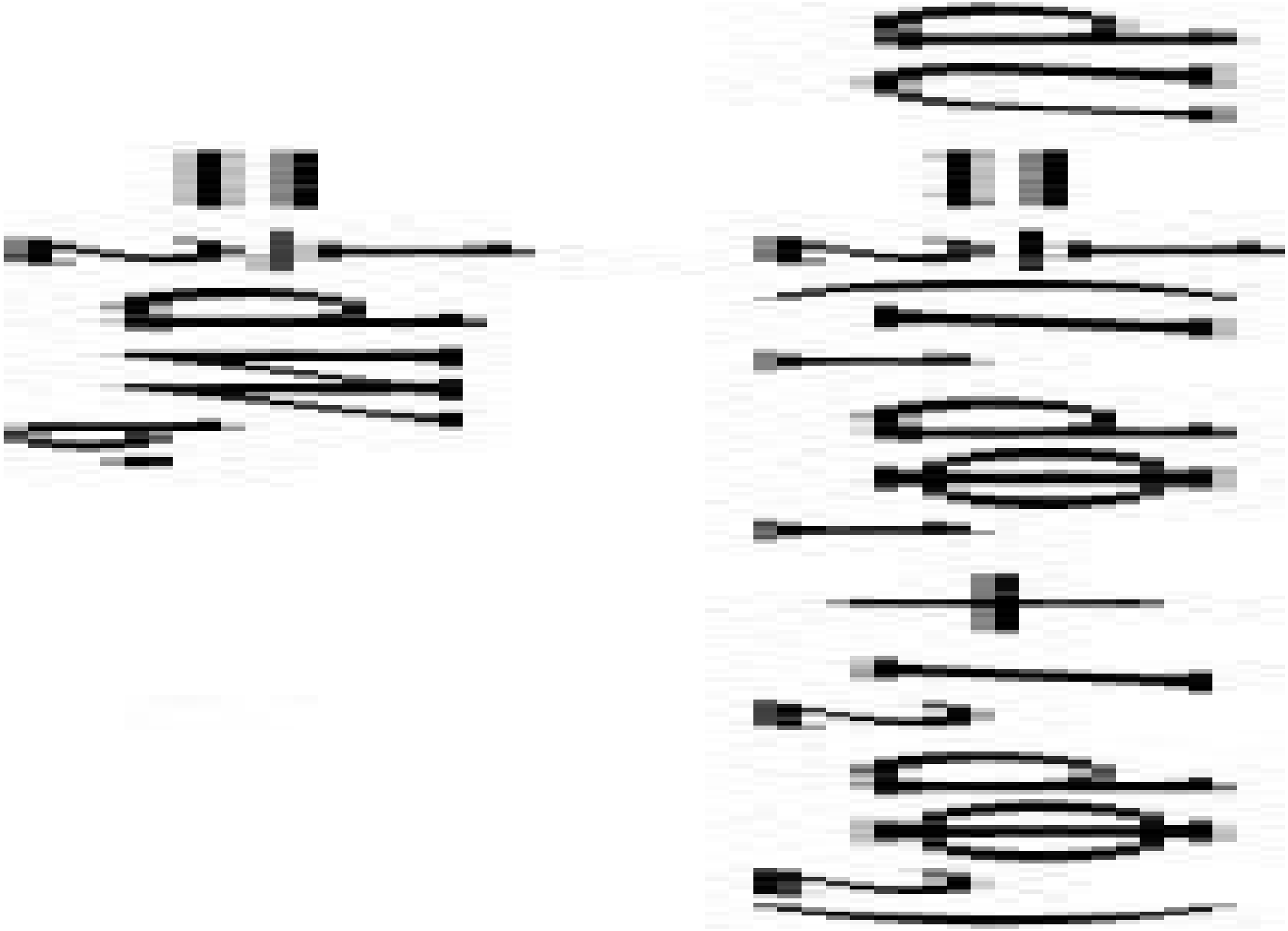
The force F is given by



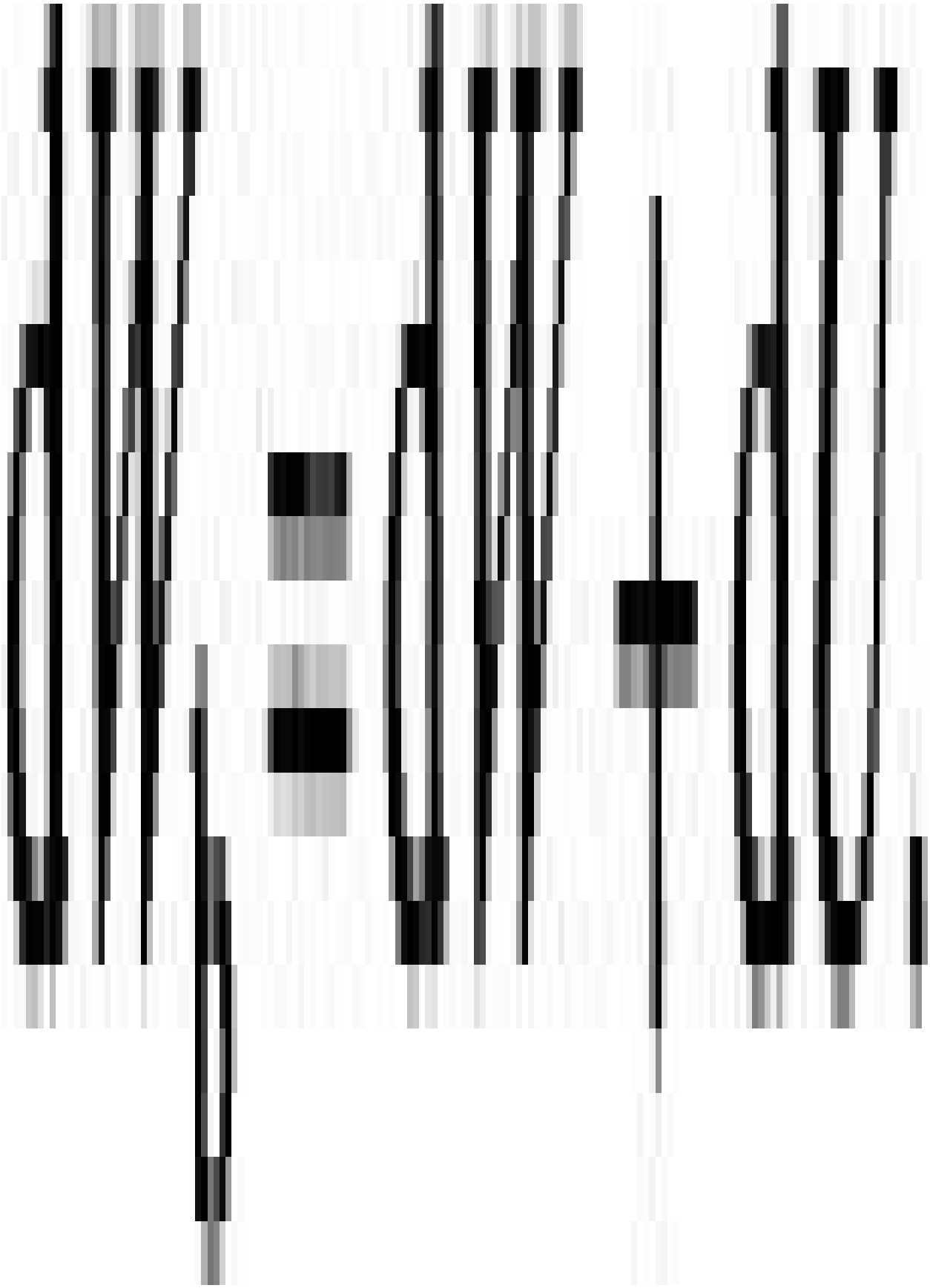
as given in section 4.6.1. During the displacement dr the batteries supplying the currents I1 and I2 do work dWb against the induced e.m.f.s in order to maintain the currents constant.



where $d\Phi 1$ and $d\Phi 2$ are the changes in flux through the circuits during the displacement dr. This result follows from Equation (6.26). If the currents are constant the change in the total magnetic energy of the system is obtained from Equation (6.27),



The work done by the batteries against the induced e.m.f.s equals the sum of the mechanical work done by the magnetic force in moving the coil and the increase in total magnetic energy, therefore



Eliminating dWb from the last two equations we obtain

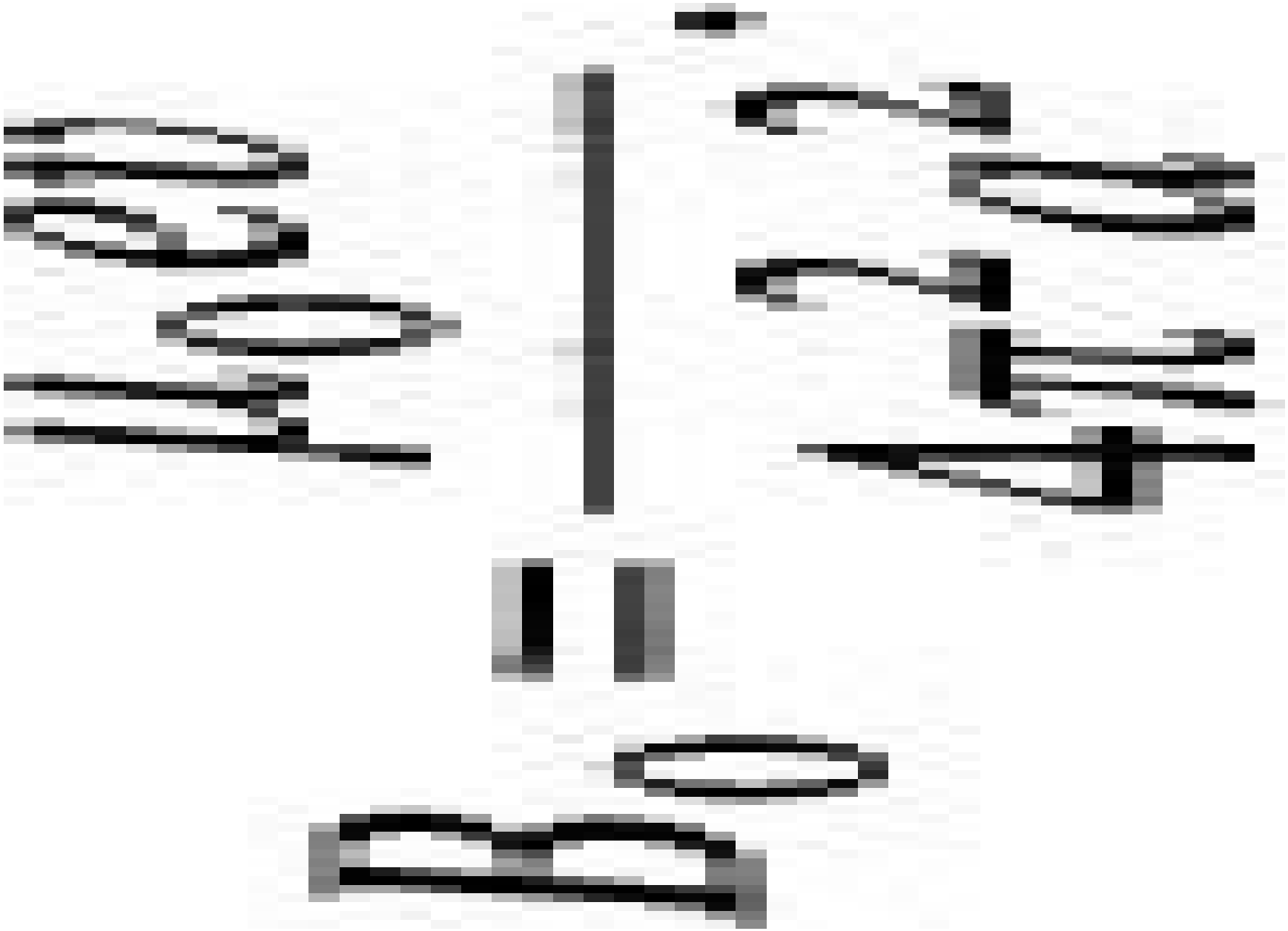
$$dU = dW$$

$$= -dU_p$$

from Equation (6.32)*. In terms of the total magnetic energy U the force on the coil F is now given by

(6.33)

This equation is very useful for determining forces on conductors or on magnetized materials in magnetic fields, for which it is also valid. For example the force between the pole faces of an electromagnet can be calculated by determining how the total magnetic energy varies with the separation of the pole faces, keeping the current in the magnetizing coils constant. Some examples are given in the problems at the end of the chapter. As a simple example of the application of ideas developed so far in this section we will obtain a rough estimate of the energies stored in the electric and magnetic fields existing in a hydrogen atom. This will give an idea of the relative importance of the electric and magnetic effects in determining atomic structure. We will again use a classical model of the hydrogen atom in which an electron is revolving around a proton in an orbit of radius a with speed v. Assuming that the electric field is little different from that due to two stationary charges, Equation (1.36) gives the value of about 8 eV for the energy stored in the electric field if we put a = 10−10 m. In order to estimate the energy U stored in the magnetic field of the moving electron, let us assume that the flux through the electron's orbit is equal to $\pi a^2 B0$, where B0 is the field at the centre due to the current $ev/2\pi a$ flowing in the circle of radius a. This field is given by Equation (4.37) to be



Hence

$$U = \frac{4\varepsilon^2 v^2}{16\pi^2 r^4}$$

from Equation (6.27). Taking the angular momentum mQva of the electron equal to the typical value of Planck’s constant h divided by 2n, we obtain

$$U \sim -\frac{\mu_0 e^2 \hbar^2}{64\pi^4 m_e^2 a^3}$$

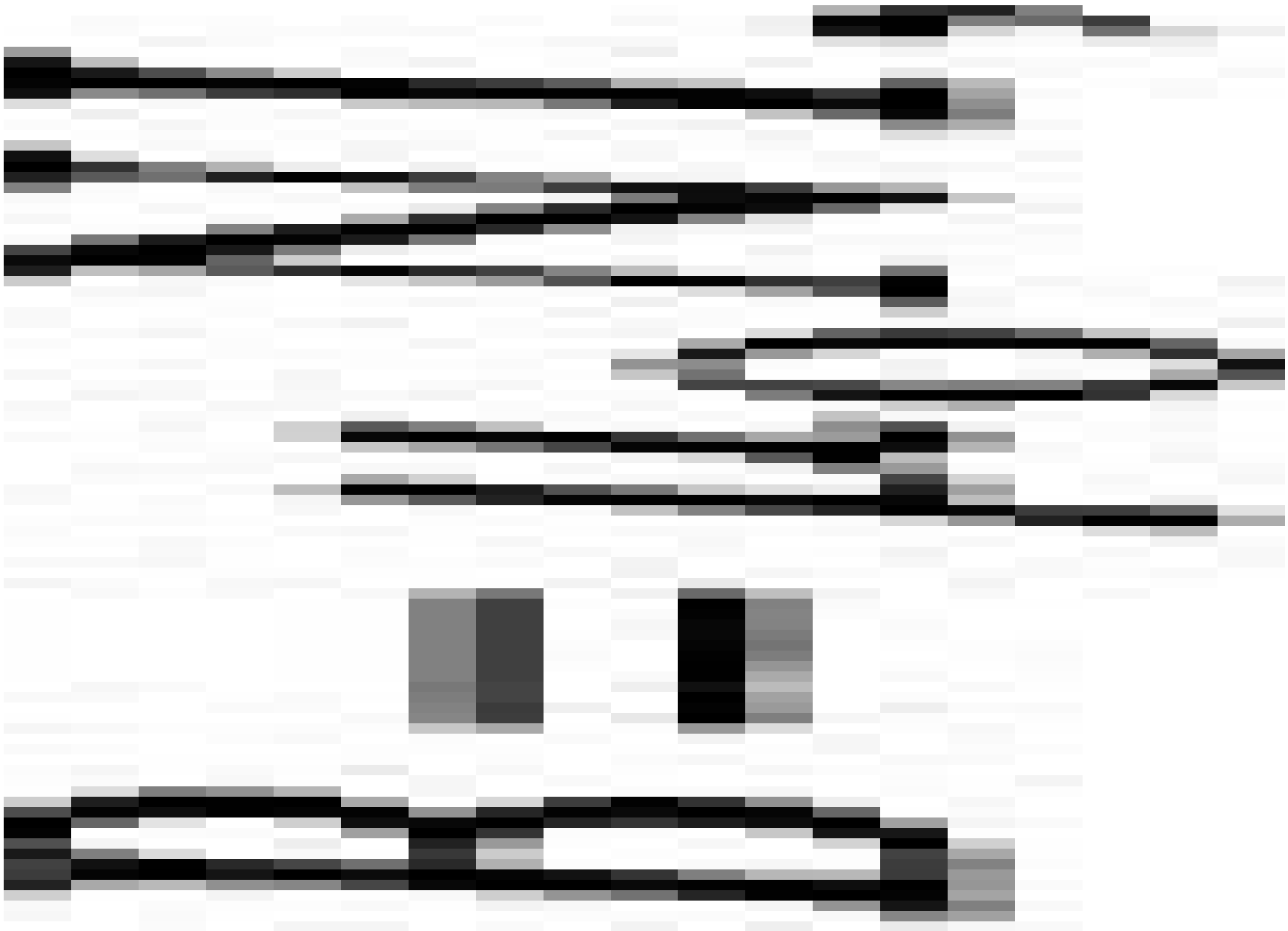
This expression has a value of about 2×10^{-5}eV. Even allowing for the inadequacies in our atomic model and in the assumptions made*, one would suspect that magnetic interactions play a part only in the finer details of atomic structure. This tentative conclusion is confirmed by proper treatments of atoms.

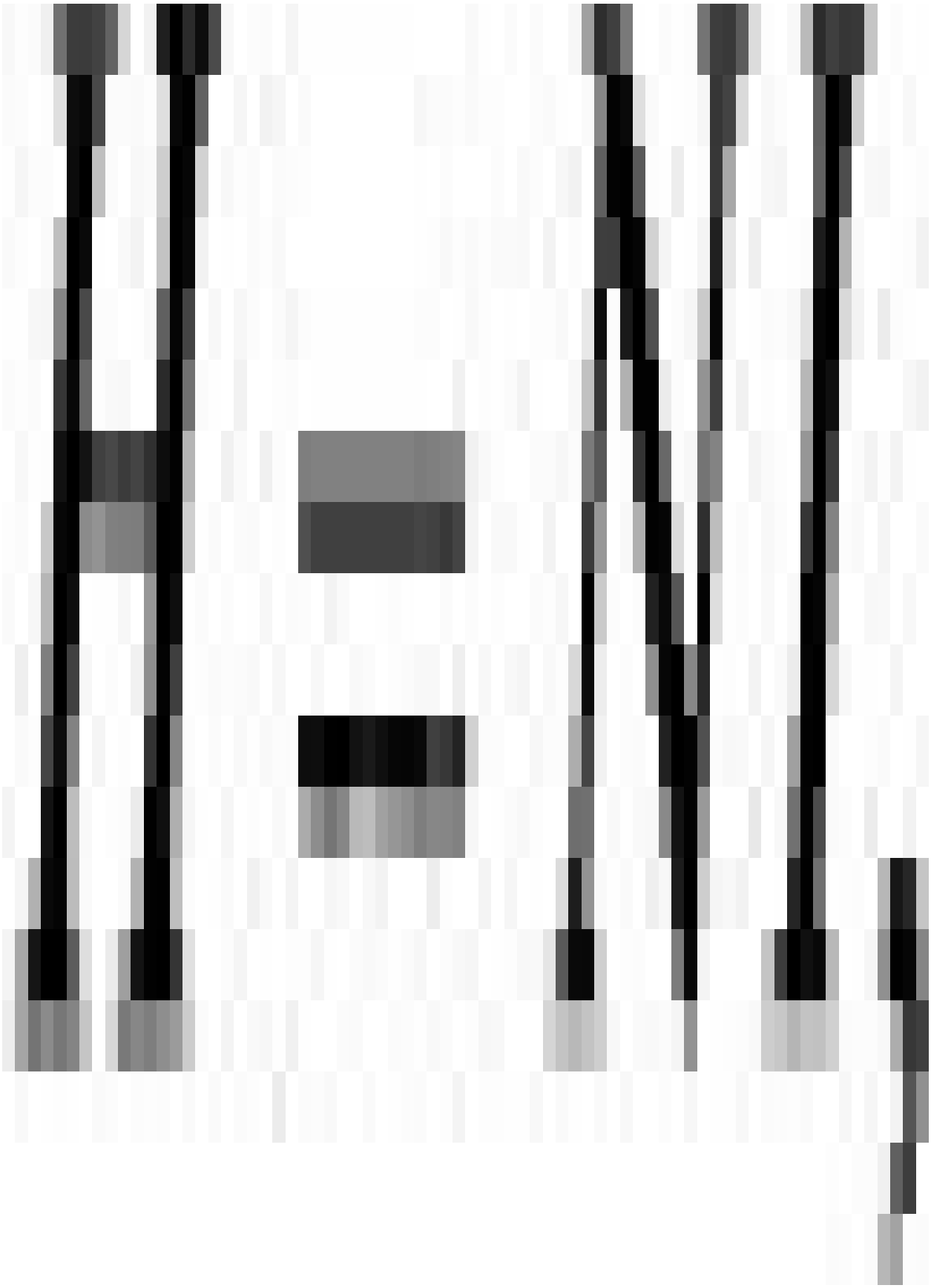
6.3.4 The total magnetic energy in terms of the fields B and H

We now return to the aim of expressing the total magnetic energy in terms of the magnetic fields B and H. We apply Equation (6.27) to the simple example of a long solenoid which has N turns per unit length, each of radius r with a current I flowing. If the solenoid has a length l the energy stored in the magnetic field is

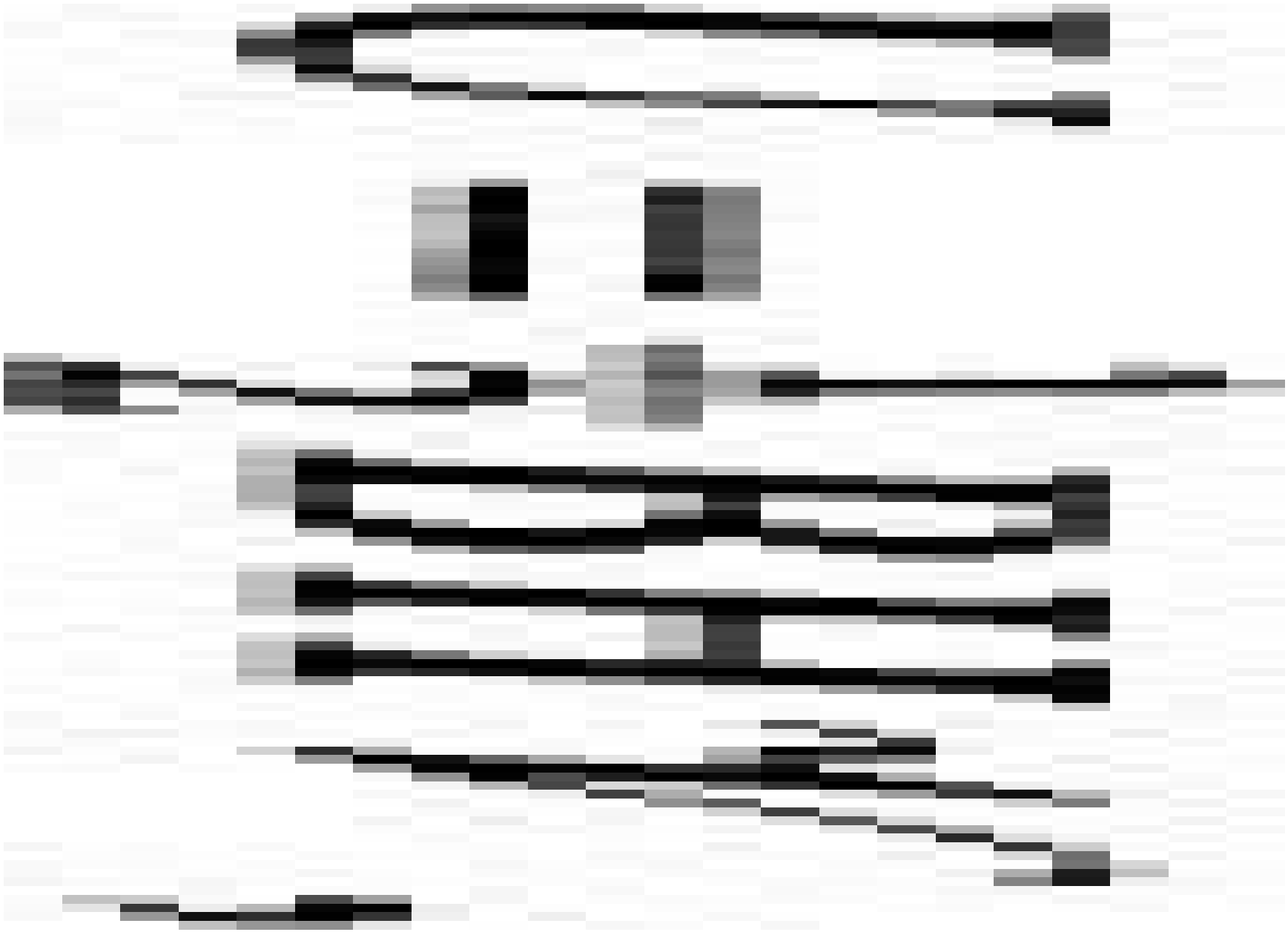
(6.34)

from Equation (6.19). The fields B and H inside the solenoid (continuing to neglect end effects) are given by

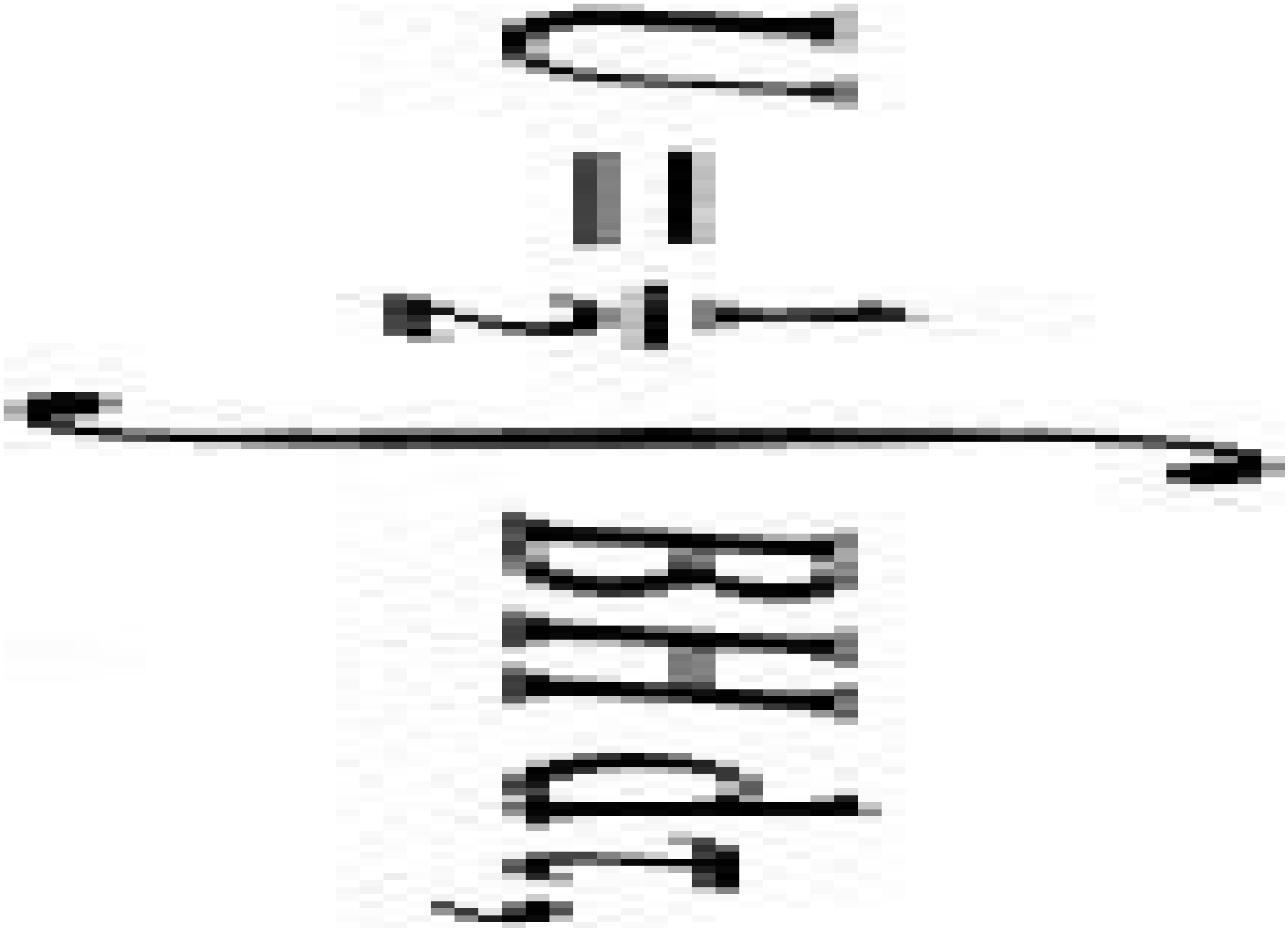




and Equation (6.34) can be written



where is the volume of the solenoid. We have neglected end effects and within this approximation the volume is the only region of space in which the fields are non-zero. The above equation for the total energy U may thus be written



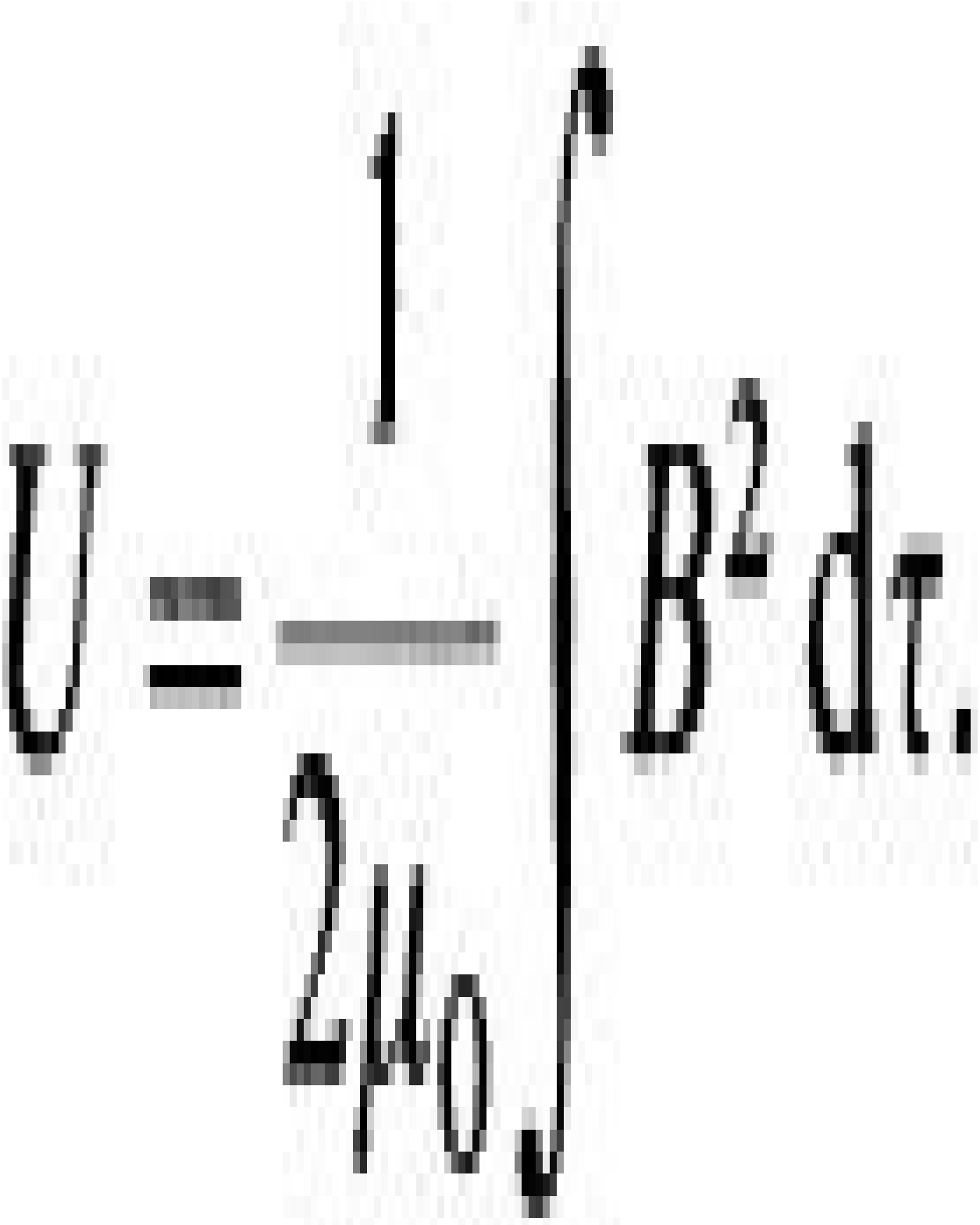
where the integral is taken over all space. The fields B and H are parallel, and so the total energy may finally be expressed in the form

(6.35)

This equation is generally true for linear media, and is the expression we would have anticipated, by analogy with Equation (2.26) for the energy stored in an electrostatic field. It is not true in the presence of non-linear media, i.e. in those situations in which the magnetic flux is not proportional to the currents in circuits. This is because its derivation depends on Equation (6.27) which itself depends on the assumption of linearity.

The energy stored in a magnetic field may be considered to be distributed with an energy density B·H. The total energy is stored in the whole field and the integral in Equation (6.35) will normally be taken over all space.

If the relative permeability can be taken to be everywhere unity Equation (6.35) becomes



Superconducting magnets can produce fields of 10 T. The energy density in such fields is about 4×10^7 joules m−3. A typical superconducting magnet might have an energy of 5×10^5 joules stored in its field, enough to boil 200 g of water. The energy density in the largest electric fields in free space, about 10^7Vm−1, is around 450 joules m−3, much less than that in large magnetic fields.

A proof of Equation (6.35) which has general validity will be giVen m section 6.3.6.

6.3.5 Non-linear media

Equation (6.35) is not valid if non-linear media like iron are present. If they are present the magnetic field B is not proportional to the currents which give rise to the field. Equation (6.26) however remains valid for that part of the work done (in changing the magnetic fluxes through the circuits) which goes to increase the stored magnetic energy.

Consider the simple situation in which a circuit is wound over a piece of magnetic material. Let the circuit have negligible resistance so that heating losses can be ignored. The work dWb done by a battery to increase the flux through the circuit by an amount $d\Phi$ in time dt is given by Equation (6.26).

(6.36)

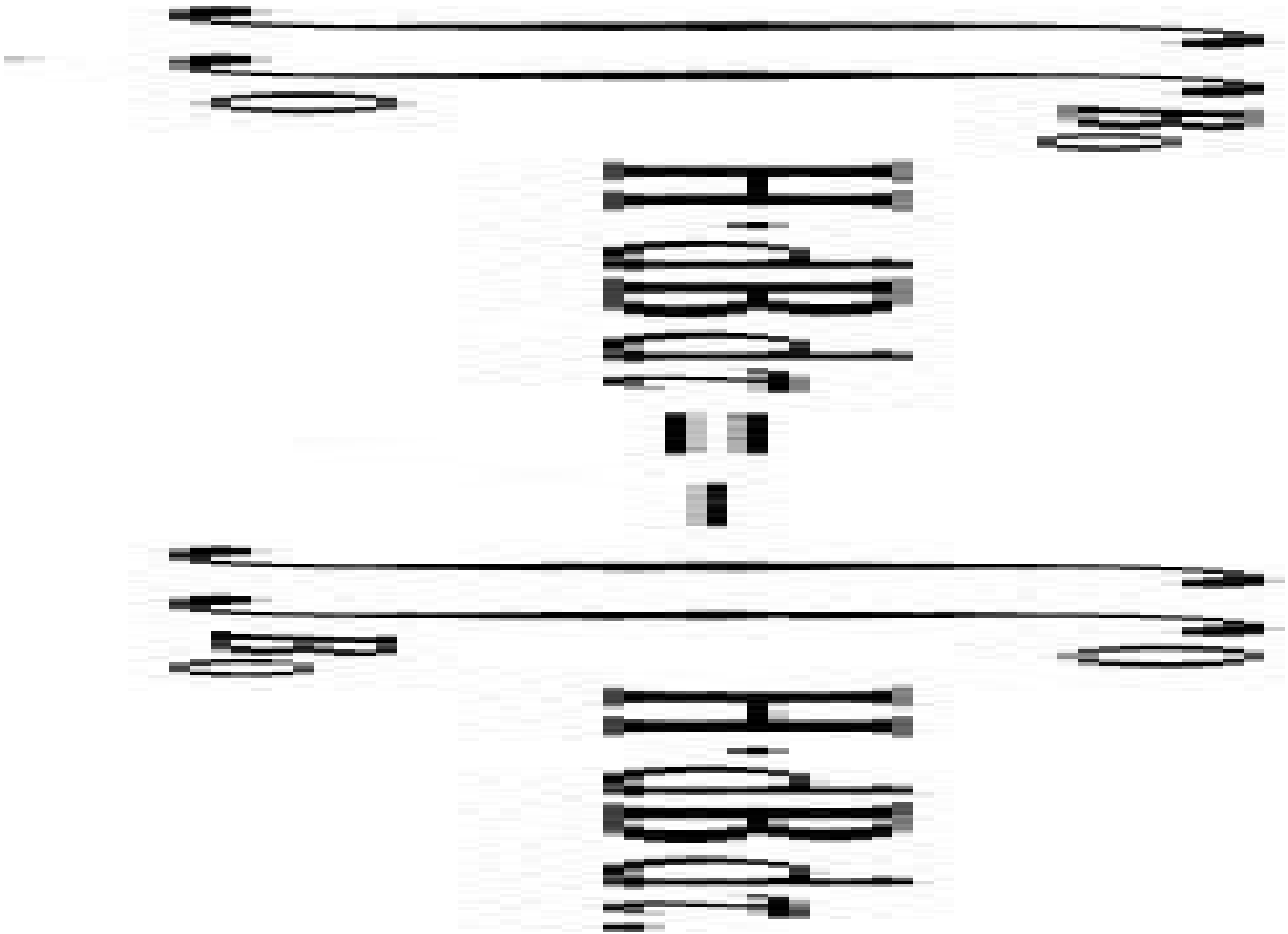
where I is the current in the circuit. By writing the flux Φ as $\int_S B \cdot dS$, where S is any surface whose perimeter is the circuit it can be shown that

(6.37)

where the volume integral is taken over all space. (The proof of this equation is given in section 6.3.6.) If the relation between the fields B and H is known, Equation (6.37) can be integrated to obtain the total work done Wb in establishing a final field B0.

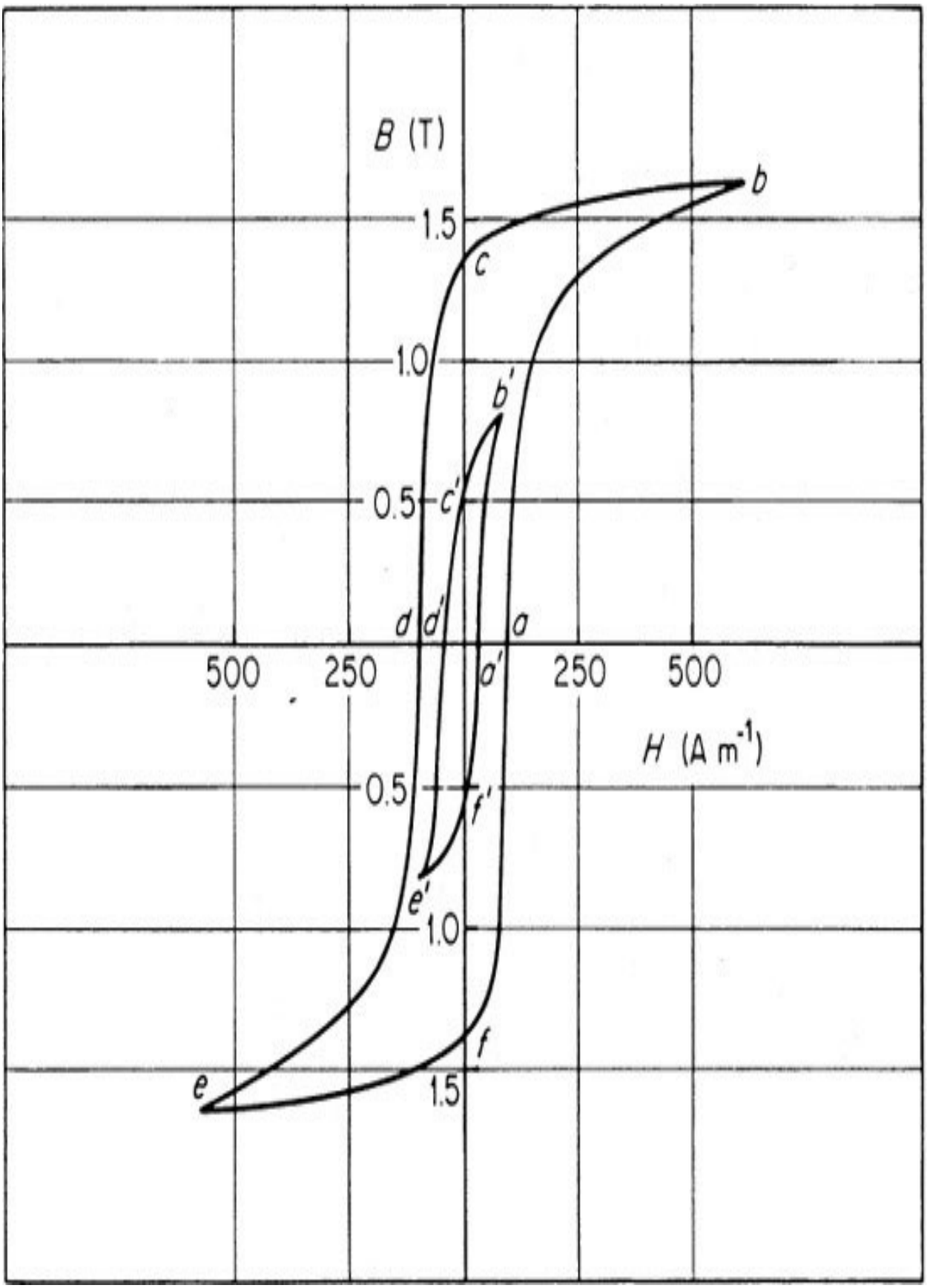
(6.38)

When the field is reduced to zero again, the total work done is zero if the relation between the fields B and H is single valued, since then



In such a situation, as we discussed in the previous section, the work done against induced e.m.f.s whilst the field is being established can be considered to be stored in the field reversibly; the energy can be regained when the field is returned to zero. However, with a piece of ferromagnetic material like iron inside the circuit, the work done to establish the field cannot all be regained, and some energy is dissipated as heat in the iron. If the iron is taken around a complete hysteresis cycle (i.e. the current through the circuit, initially at a certain value, is decreased through zero to the same value in the opposite direction and then returned to its original condition), the energy dissipated as heat in unit volume of the iron can be related to the area under the hysteresis curve*. The curve labelled abcdefa on Figure 6.10 is the major hysteresis curve for commercial iron. If unit volume of commercial iron were taken around this cycle the current sources have to do an amount of work equal to the integral around the cycle. This energy is about 350 joules m−3 as can be estimated from the figure. Iron cores are used in transformers in order to obtain maximum flux linkage between primary and secondary windings. The fields in a mains transformer core are taken through a hysteresis cycle 50 times every second. Transformers are not used under conditions where the iron is near saturation. A typical situation is one in which the fields follow the minor hysteresis loop labelled a'b'c'd'e'f'a' on Figure 6.10, and the area under this loop is about one sixth of the area under the major hysteresis loop. If 1000 cm^3 of commercial iron is used in a transformer which has currents flowing at 50 Hz, the power dissipation is about (350 ×) × 50 × 10−3 W, i.e. about 3 watts. Hysteresis losses are proportional to frequency, since current sources haveto do an amount of work per unit volume of material every cycle of the current.

Figure 6.10. The major hysteresis curve for commercial iron. The minor hysteresis curve shown is a typical cycle around which the iron in a mains transformer is taken.


B (T)
1.5
1.0
0.5
0.5
1.0
1.5
500
250
250
500
H (A m-1)
a
b
c
d
e
f
a'
b'
c'
d'
e'
f'

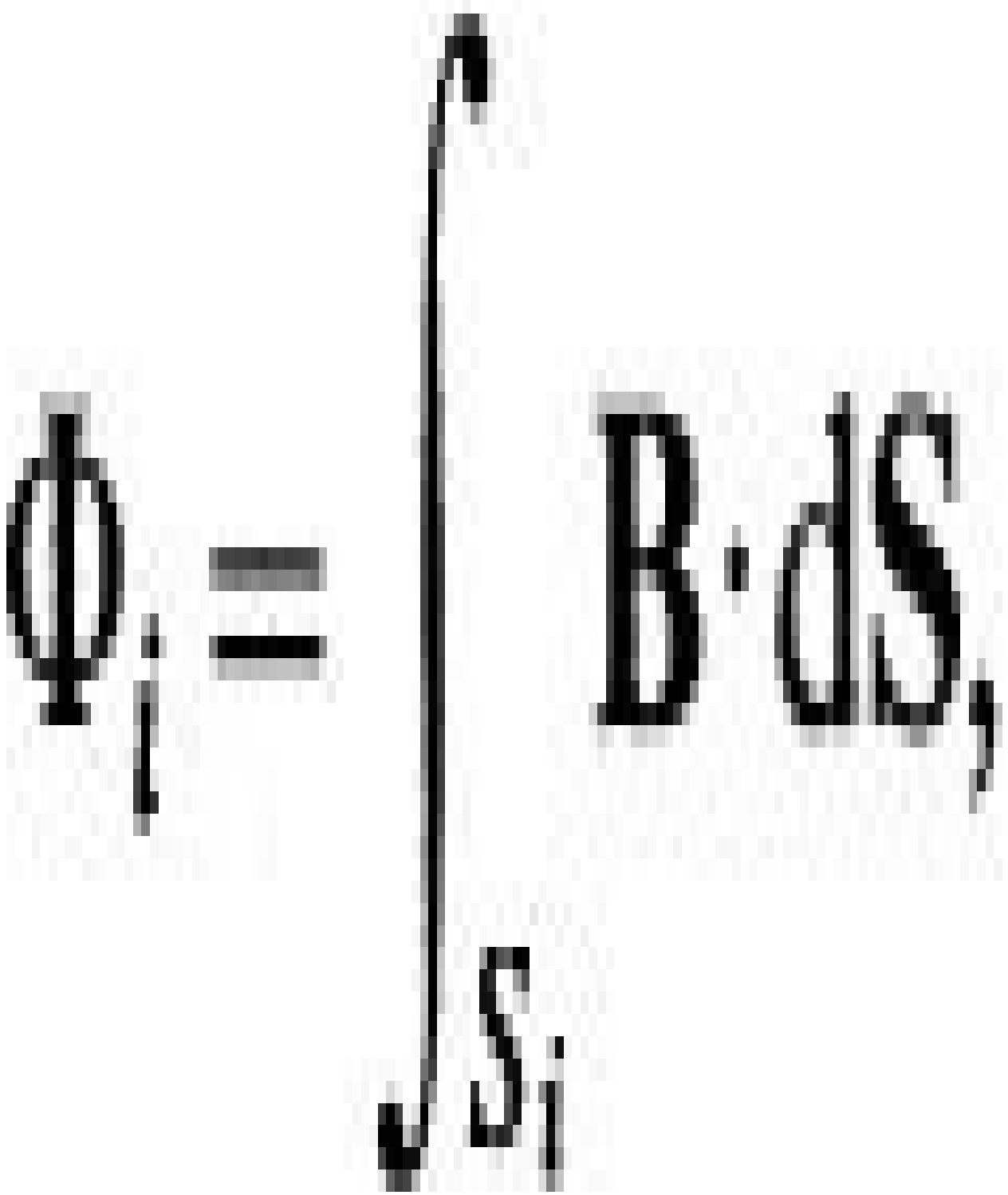
Some of the hysteresis losses in ferromagnetic materials can be attributed to the friction occurring as the domains slide past each other in reorientating with the changing field. Losses of a somewhat similar kind can occur in capacitances filled with polar dielectrics used in alternating current circuits. These losses are usually much smaller than hysteresis losses, and arise when the small fraction of the molecules reorientate with the changing electric field and suffer ‘friction’ from their neighbours in so doing.

6.3.6 Further comments on energy in magnetic fields

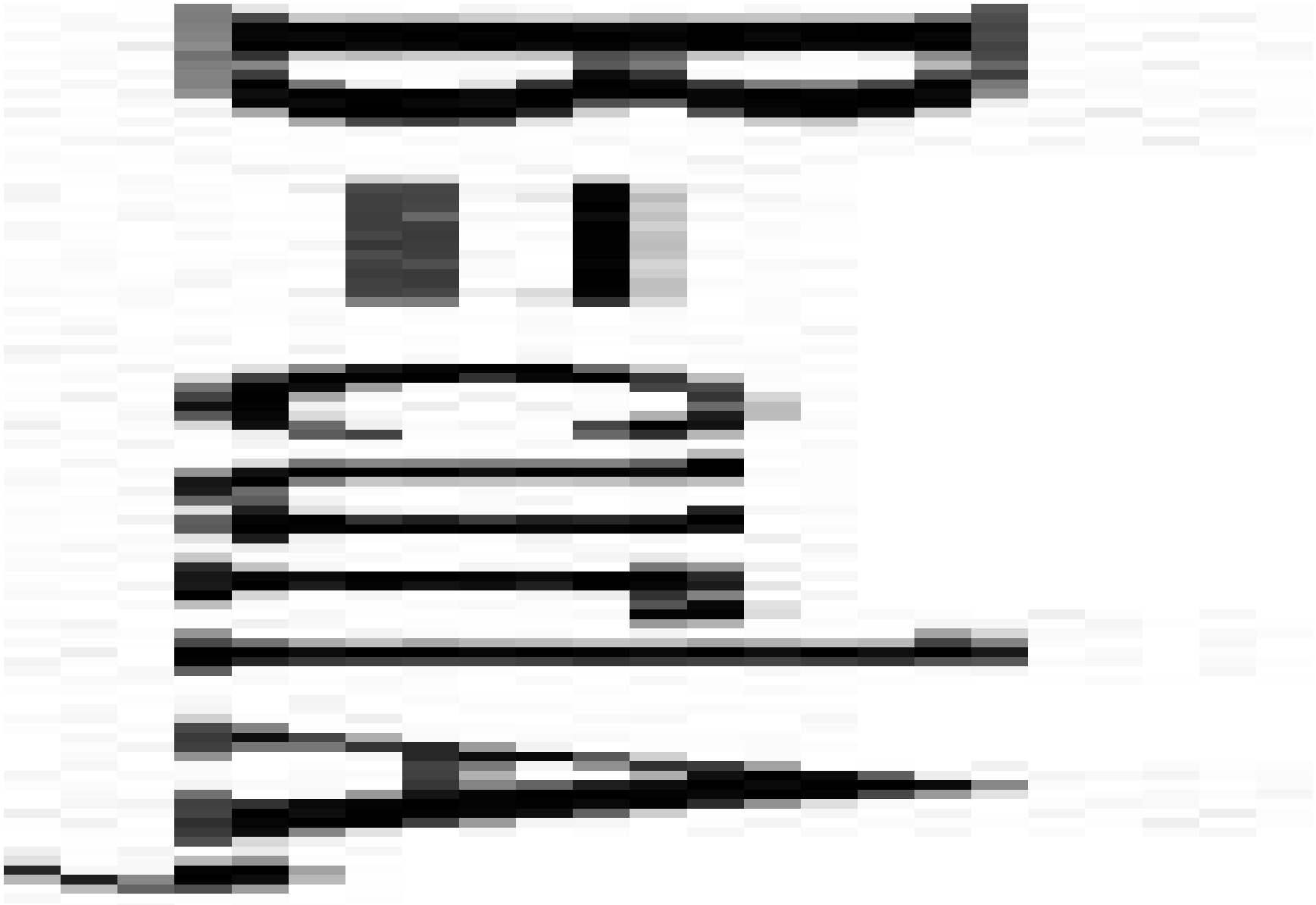
In this section, which can be omitted without loss of continuity, we give more general derivations of two important expressions discussed previously. The first is a proof of Equation (6.35) which gives the total energy stored in a magnetic field in terms of the fields B and H. The second is a proof of Equation (6.37) which gives the work done by a current source when it effects a small increase in magnetic field B in a medium.

We first prove Equation (6.35), and emphasize that for now we confine our attention to linear magnetic media. The work done in setting up a given charge distribution can be calculated with the aid of the inverse square law, and written down neatly, as in Equation (2.27), in terms of the charges and the electrostatic potential ϕ. Similarly the work done in setting up currents in inductive circuits can be calculated with the aid of Faraday’s law, and written down in terms of the currents and the magnetic vector potential A.

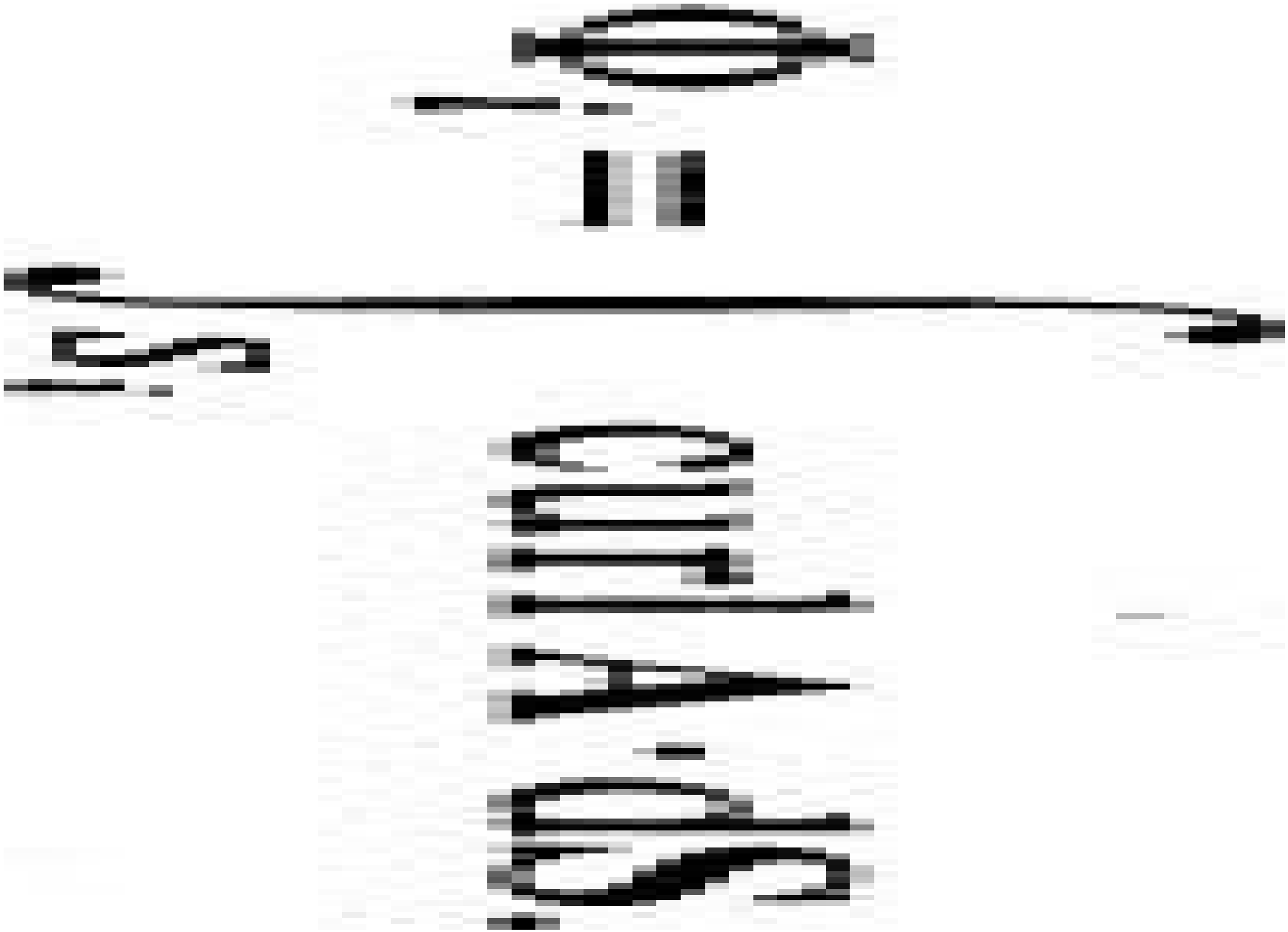
Suppose that currents Ii flow in n circuits through which are fluxes Φi. The flux Φi through the ith circuit is



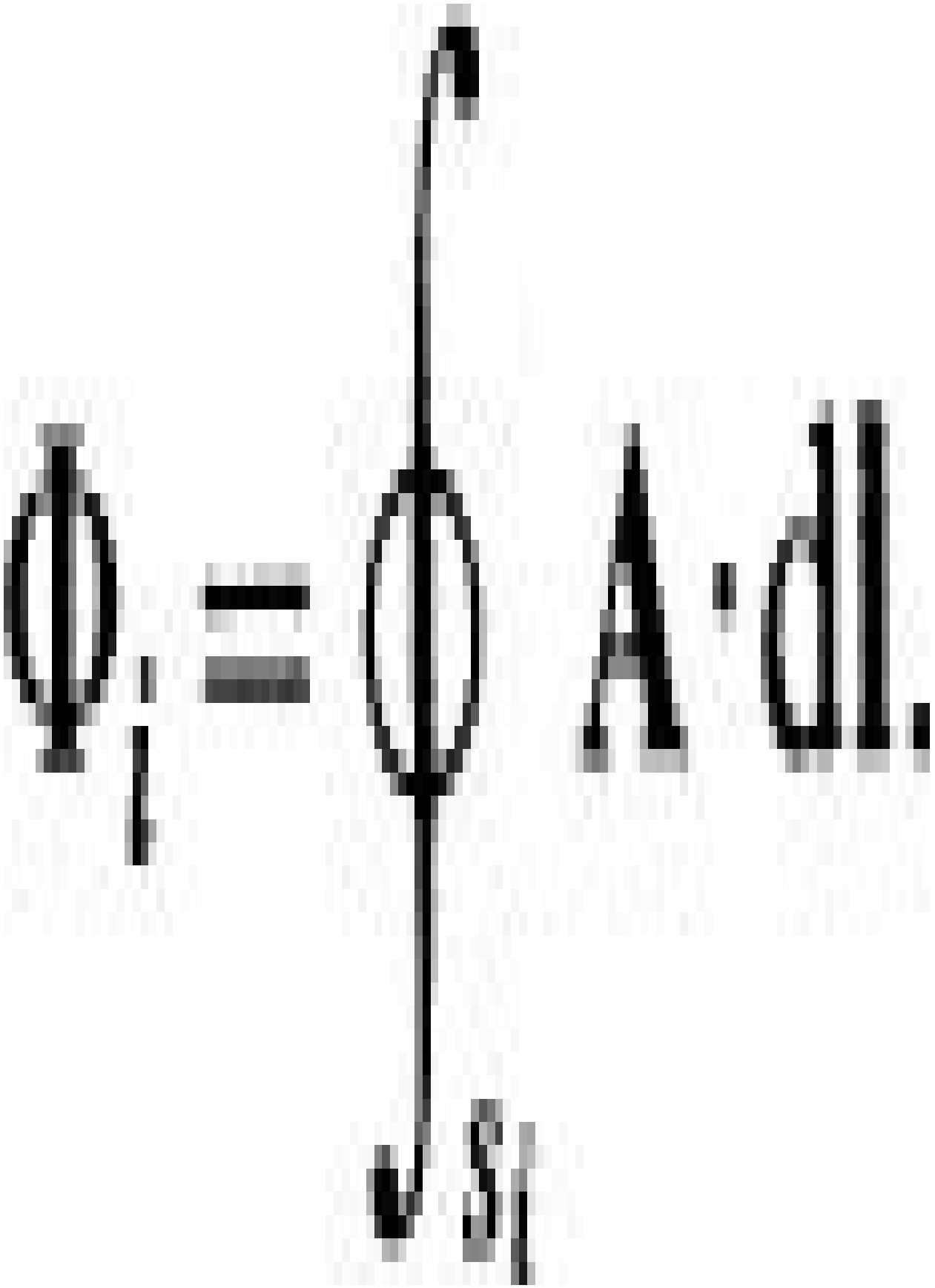
where B is the total magnetic field. In terms of the vector potential A



and hence



The right-hand side of this equation can be transformed to a line integral using Stokes' theorem, to give



The total magnetic energy U is now given by Equation (6.27),

(6.39)

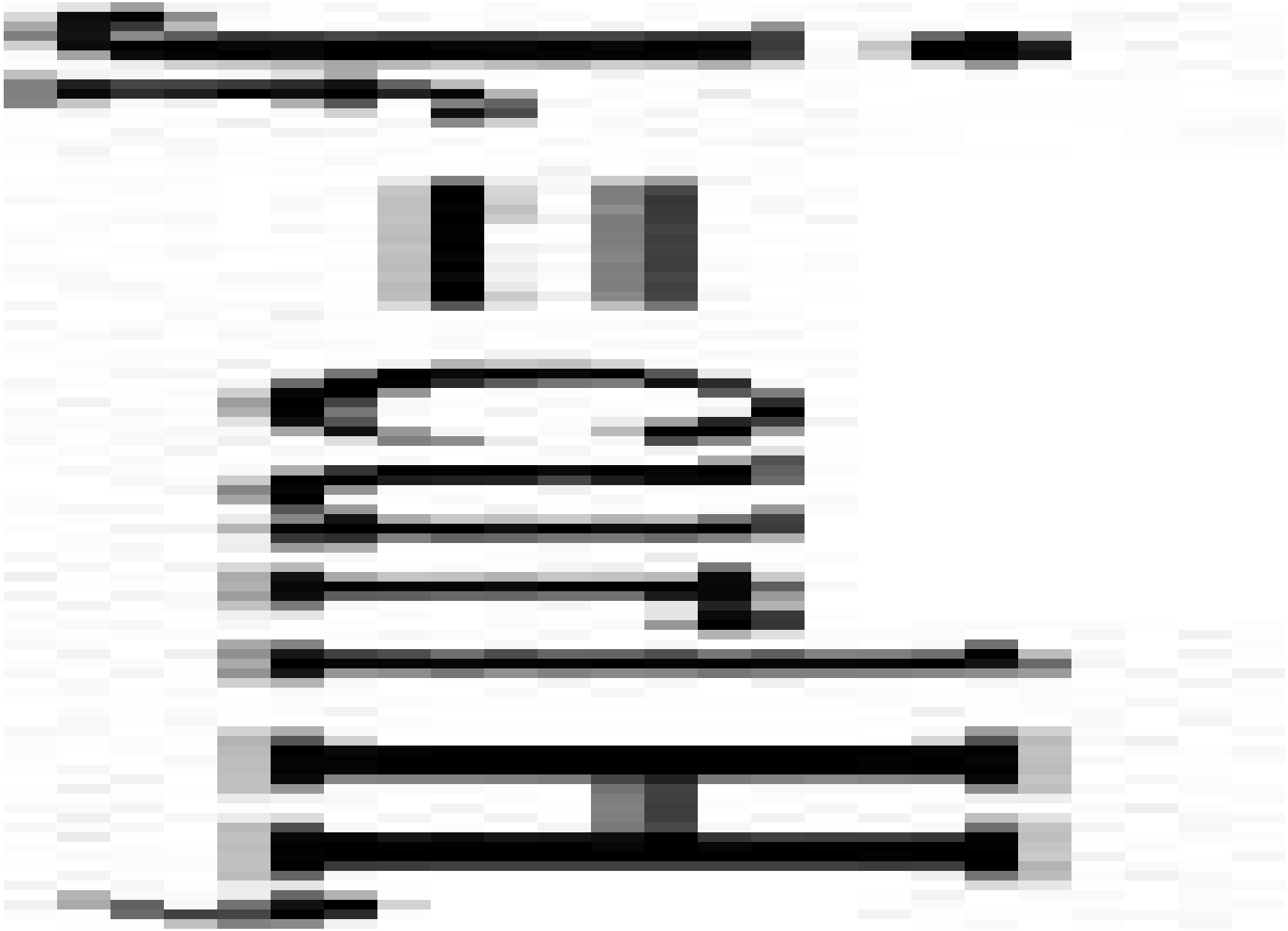
where the direction of dl is the same as that of the current Ii. As mentioned above, in the calculation of self-inductance, current-carrying wires can not be assumed to be infinitesimally thin, and integration must be made over the volume of the wire. Similarly here the quantity Ii dl should be replaced by jfdτ, where jf is the free current density and dτ the volume occupied by the length dl of the wire. Wherever there is no wire the current density is zero. The sum of line integrals in Equation (6.39) can thus be replaced by an integral over all space, with the variable Ii dl replaced by jf dτ. We obtain

(6.40)

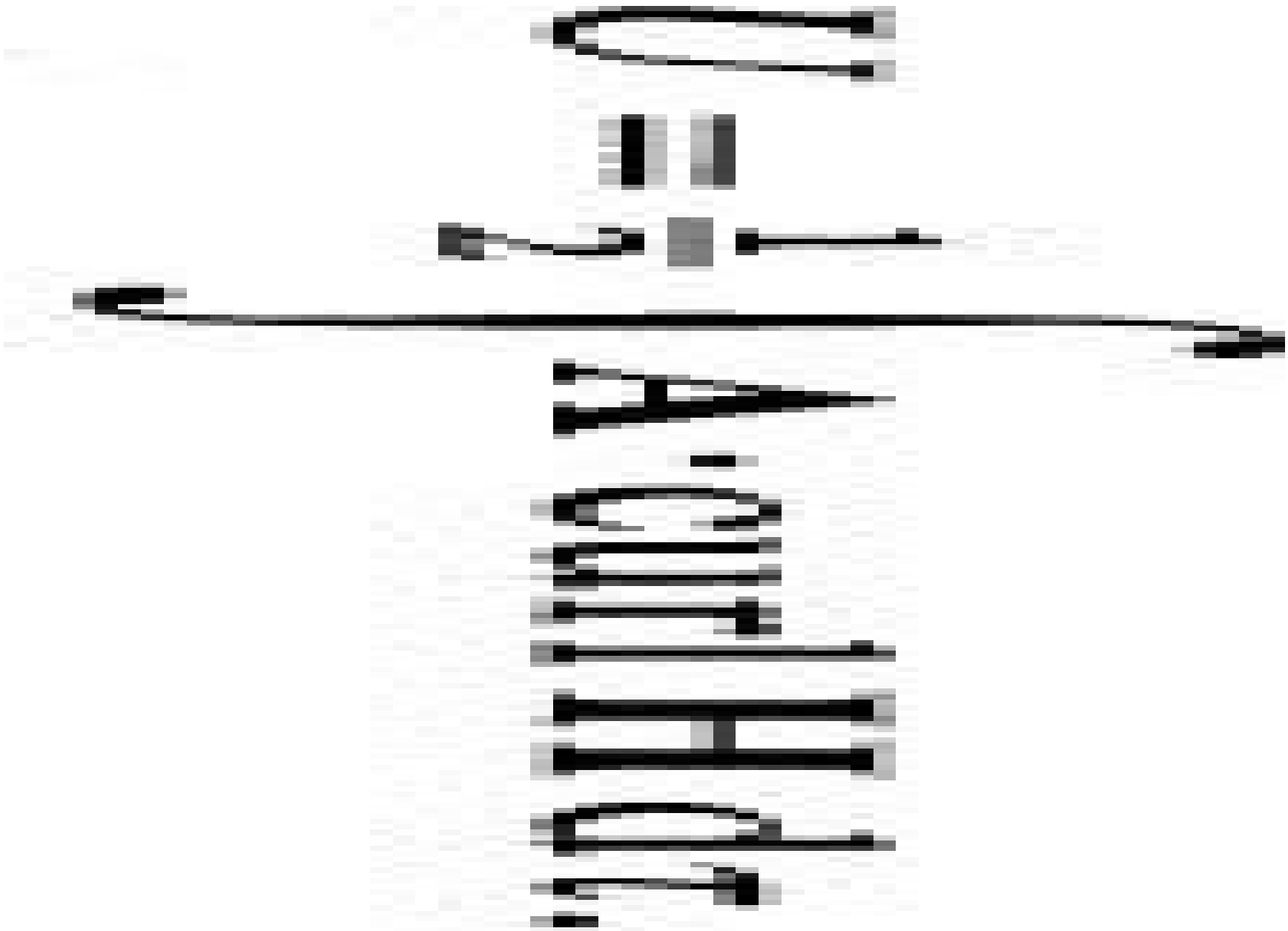
This formula should be compared with Equation (2.27) for the energy in an electrostatic field

$$V = \frac{1}{3}\pi r^2 h + \frac{2}{3}\pi r^3$$

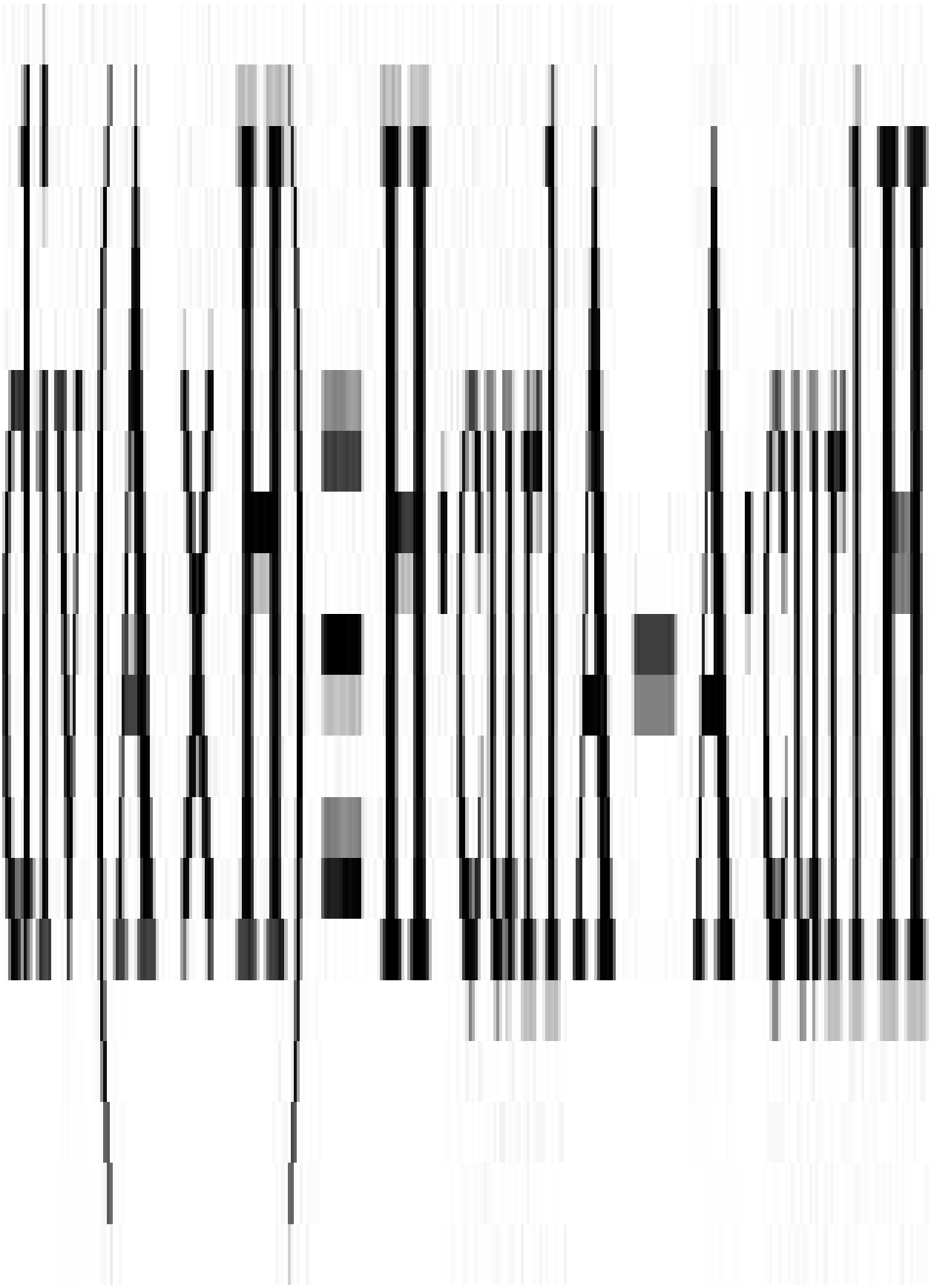
where ρf and ϕ are the electric charge density and electrostatic potential respectively, and σf is the surface density of free charge on conducting surfaces S. Equation (6.40) can be further transformed by noting that the free current density jf and the magnetic intensity H are everywhere related by Equation (5.30)



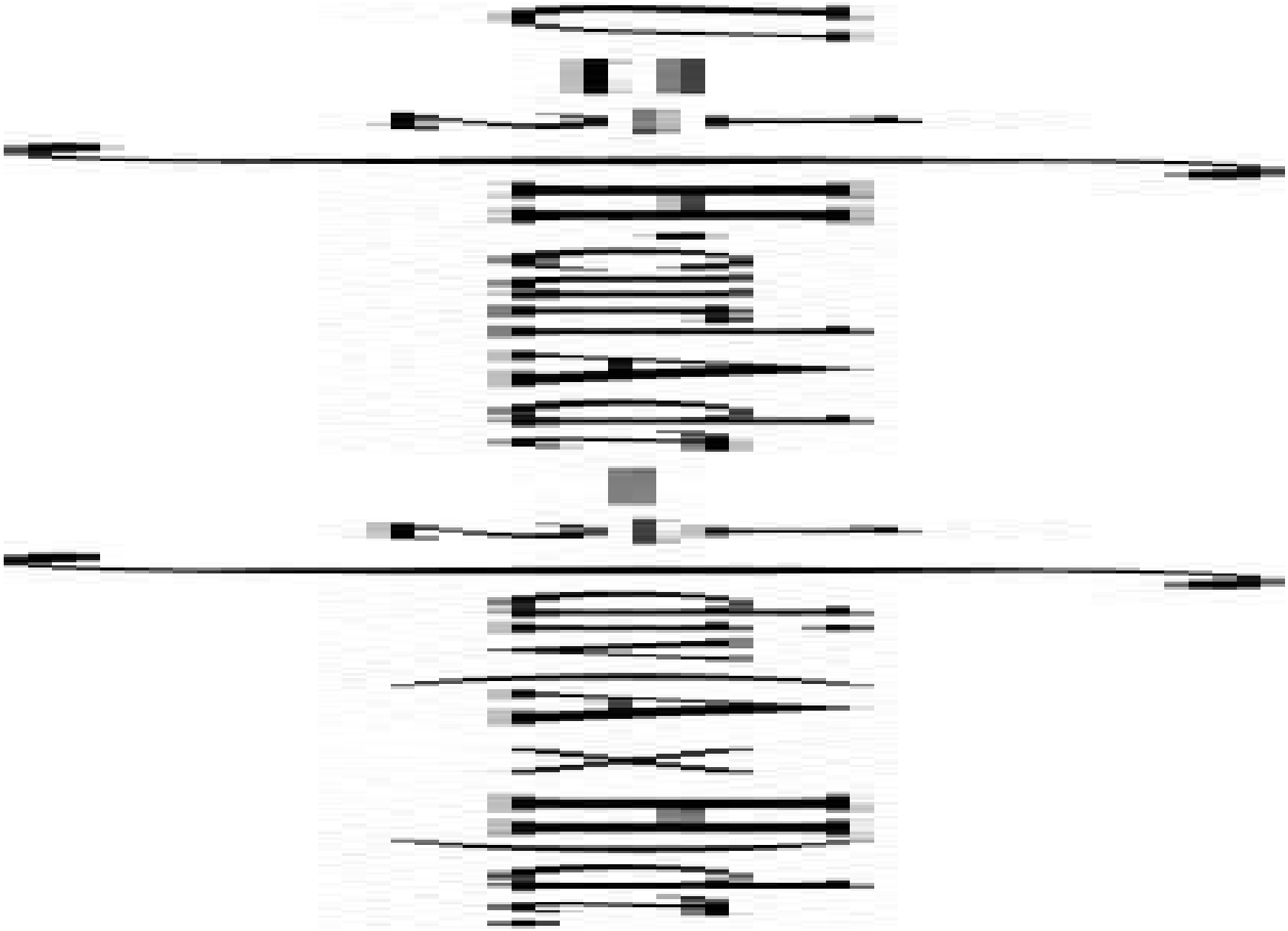
and substituting for jf in Equation (6.40)



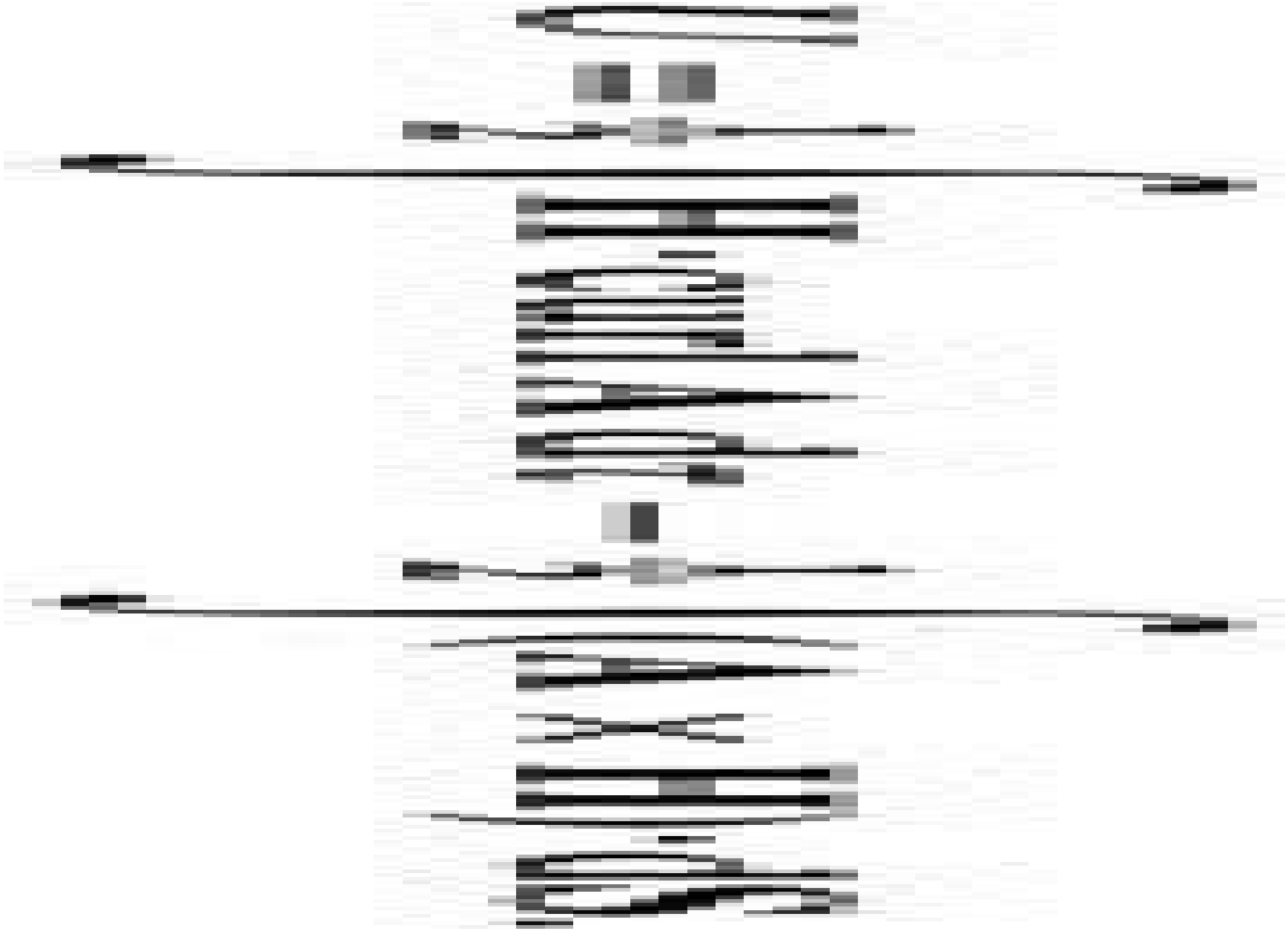
By using the vector identity (Appendix B)



the above equation becomes



The second integral on the right-hand side can be transformed into a surface integral using the divergence theorem (1.21), and



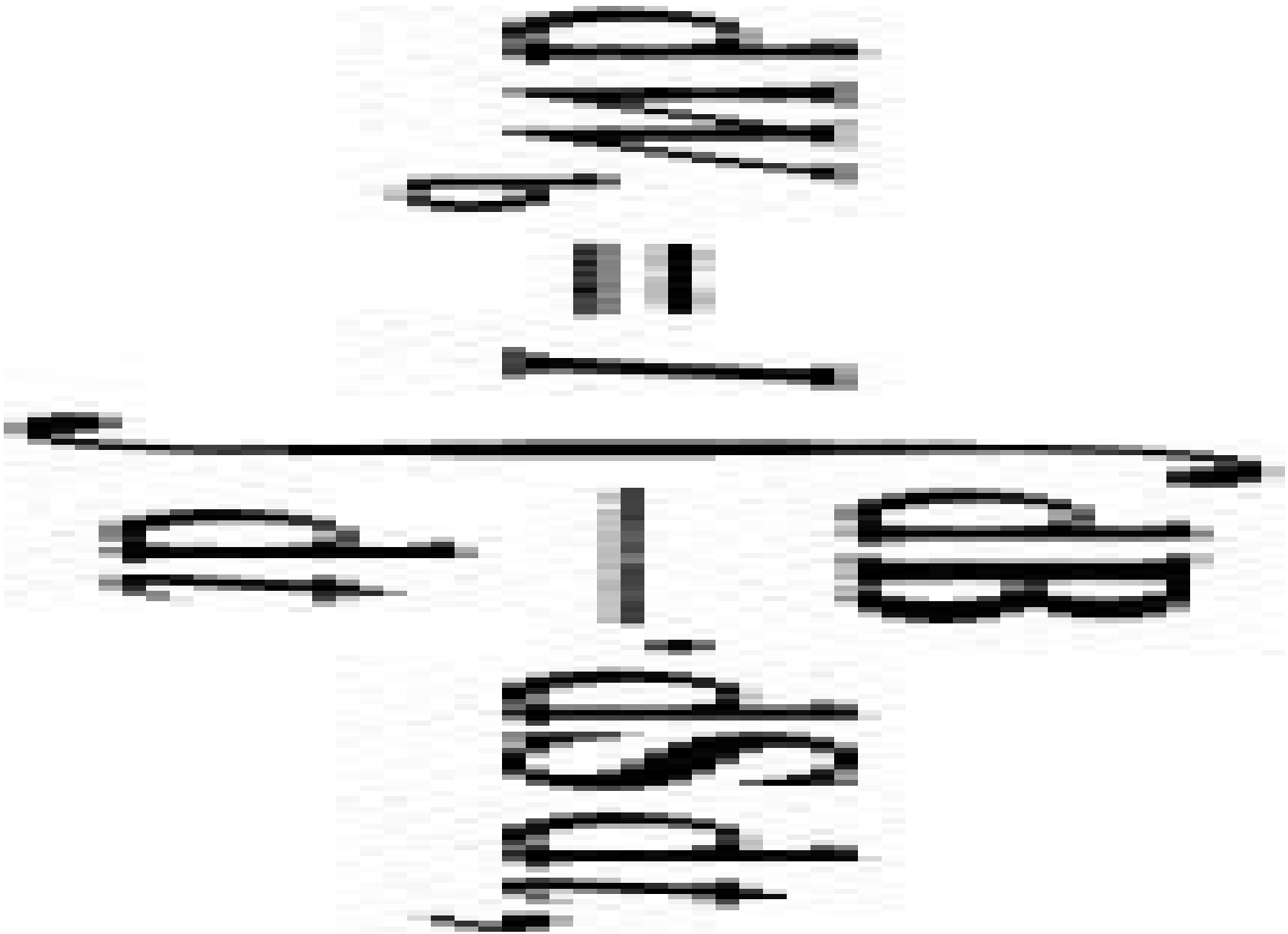
The surface integral is to be evaluated over an infinite surface bounding all space. However, we may choose a finite surface at very great distances from the sources of the field. Over this surface the fields are negligibly small, since they fall off with distance r at least as fast as r^{-2}. The vector $(A \times H)$ thus falls off at least as fast as r^{-3} and the surface integral becomes indefinitely small. We thus obtain finally

$$U = \frac{1}{2}\int \mathbf{H}\cdot\operatorname{curl}\mathbf{A}\,d\tau$$

$$= \frac{1}{2}\int \mathbf{H}\cdot\mathbf{B}\,d\tau,$$

the volume integral being taken over all space. This is the same result as Equation (6.35)

We now proceed to the proof of Equation (6.37). From Equation (6.36)



where S is a surface spanning the contour of the circuit. From Equation (6.17)

$$dW_b = -I \int_S \operatorname{curl} \mathbf{E} \cdot d\mathbf{S}\, dt$$

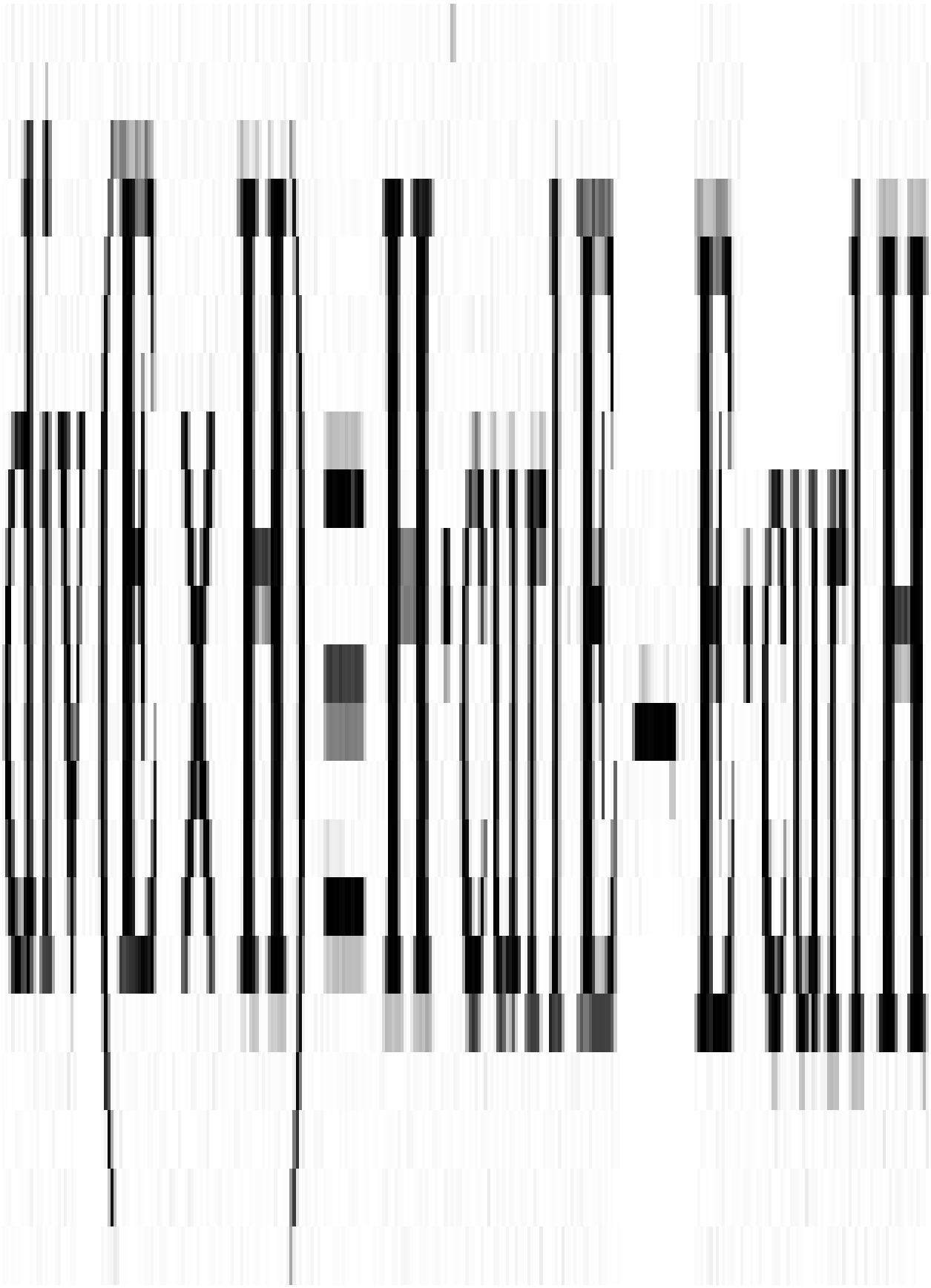
$$= -I \oint \mathbf{E} \cdot d\mathbf{l}\, dt,$$

from Stokes' theorem. If the current I is specified by a current density jf, the line integral above can be replaced by a volume integral over all space, and

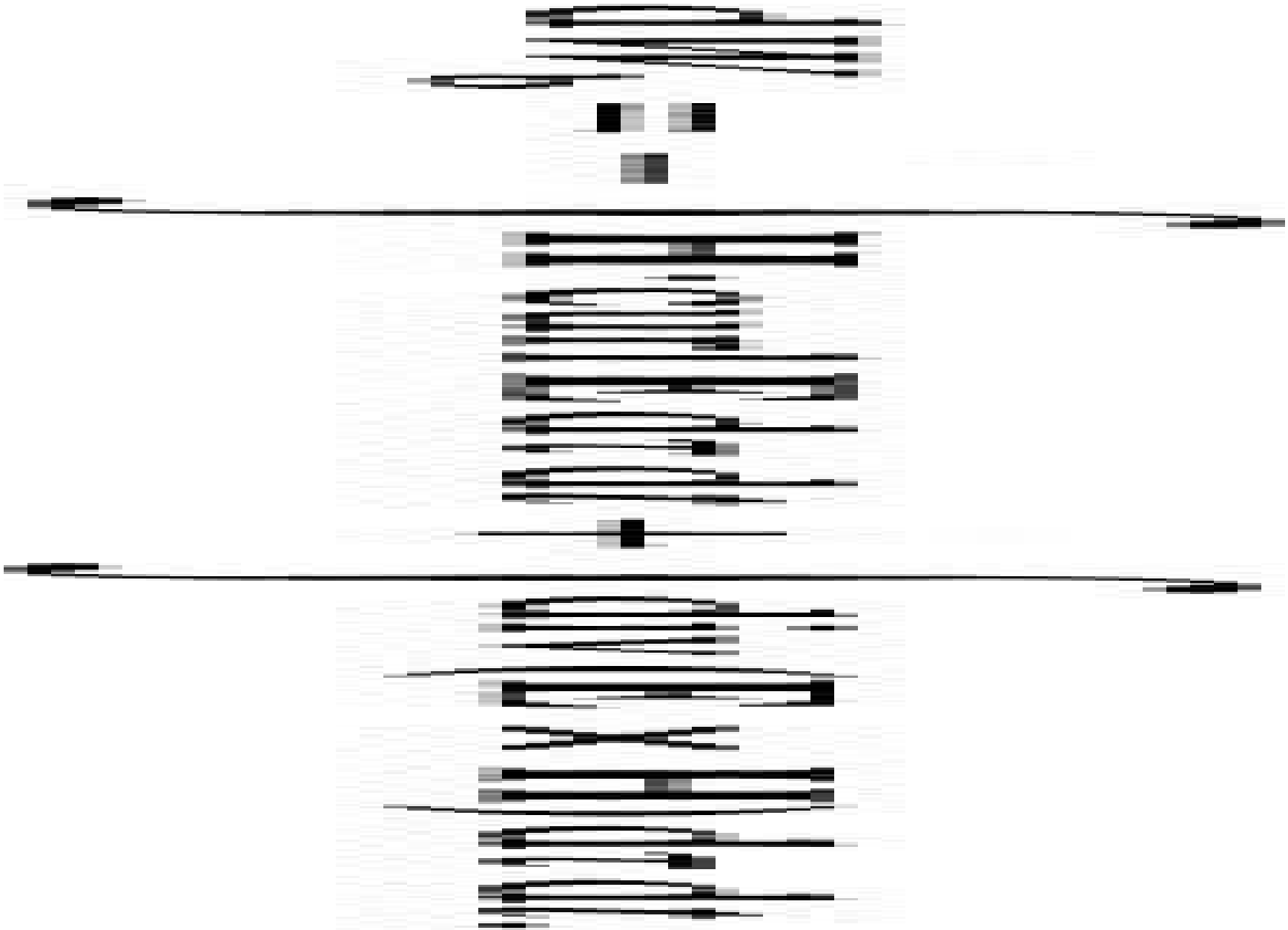
$$dW_p = -\int \mathbf{E} \cdot \mathbf{j}_f \, d\tau \, dt$$

$$= -\int \mathbf{E} \cdot \operatorname{curl} \mathbf{H} \, d\tau \, dt,$$

from Equation (5.30). Using the vector identity (Appendix B)



we obtain



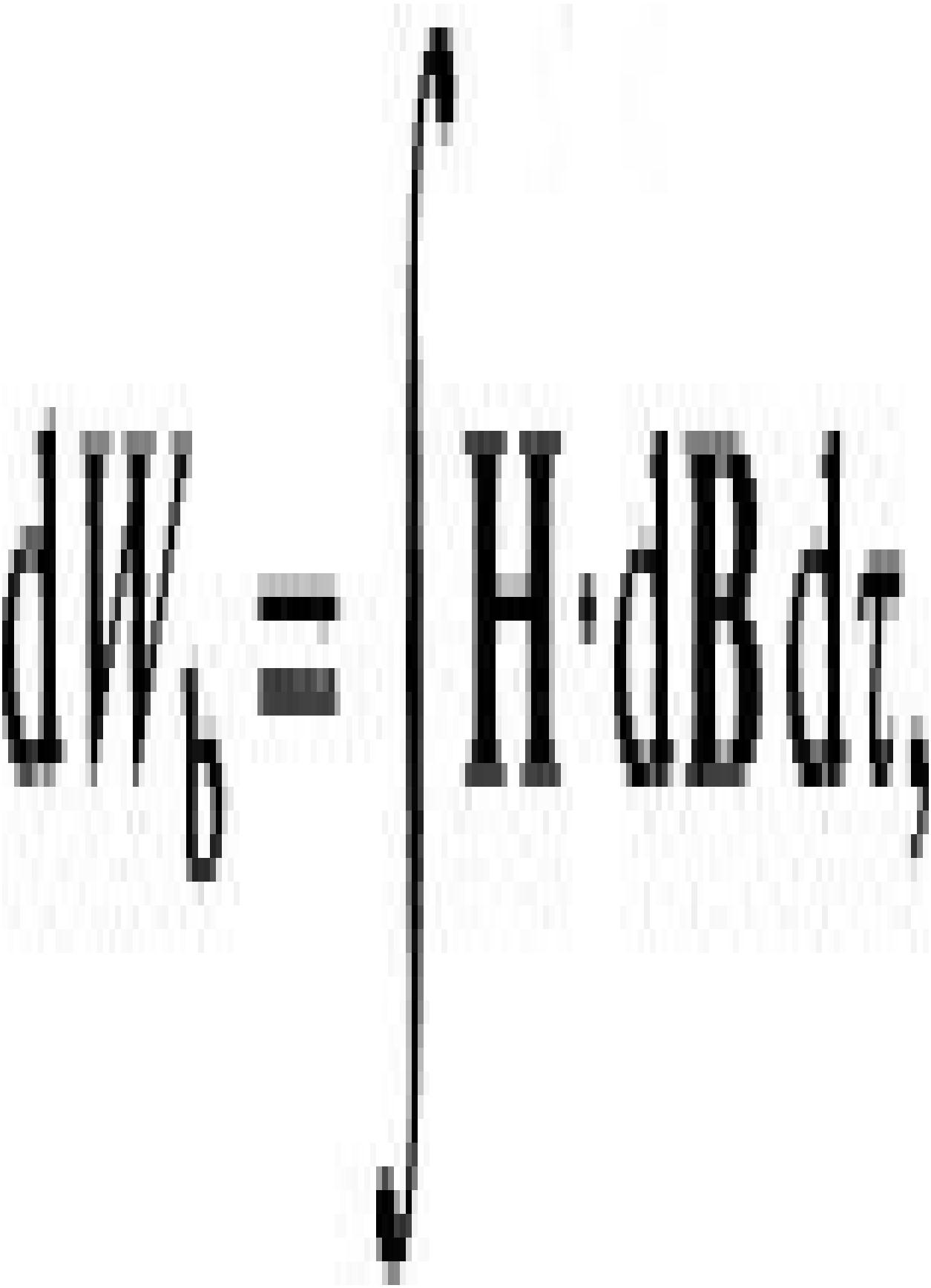
The second integral on the right-hand side of this equation can be transformed into a surface integral using the divergence theorem (1.21). Its value is zero since the fields E and H vanish over a surface a very long way from the circuit.

Hence

$$dW_b = -\int_{\downarrow}^{\uparrow} \mathbf{H} \cdot \operatorname{curl} \mathbf{E}\, d\tau\, dt$$

$$= \int_{\downarrow}^{\uparrow} \mathbf{H} \cdot \frac{\partial \mathbf{B}}{\partial t}\, d\tau\, dt$$

from Equation (6.17). We see that this is the same result as Equation (6.37),



which we used in the discussion of hysteresis.

6.4 THE MEASUREMENT OF MAGNETIC FIELDS AND SUSCEPTIBILITIES

Current can be measured by determining the force between two circuits carrying the current and relating the force and the current by the Biot–Savart law. A moving coil instrument of the kind described in section 4.6.1 can be calibrated in this way to give a convenient current meter. The measurement of the force or torque on a circuit carrying a known current in a magnetic field B provides an absolute measurement of the field. Such measurements however need sensitive equipment and are not used for everyday field determinations in the laboratory. For example a coil of area 1 cm^2 which consisted of 100 turns of thin wire could carry a current of about 1 A before overheating. If it were placed with its plane parallel to a field to 0.1 T the torque on it would be the same as that due to the weight of a mass of 10^{-7} g acting over a lever arm 1 cm long. Even if such a small torque could be accurately measured, the device would be measuring the average field over the area of the coil, and so would not be useful for fields which varied rapidly with position.

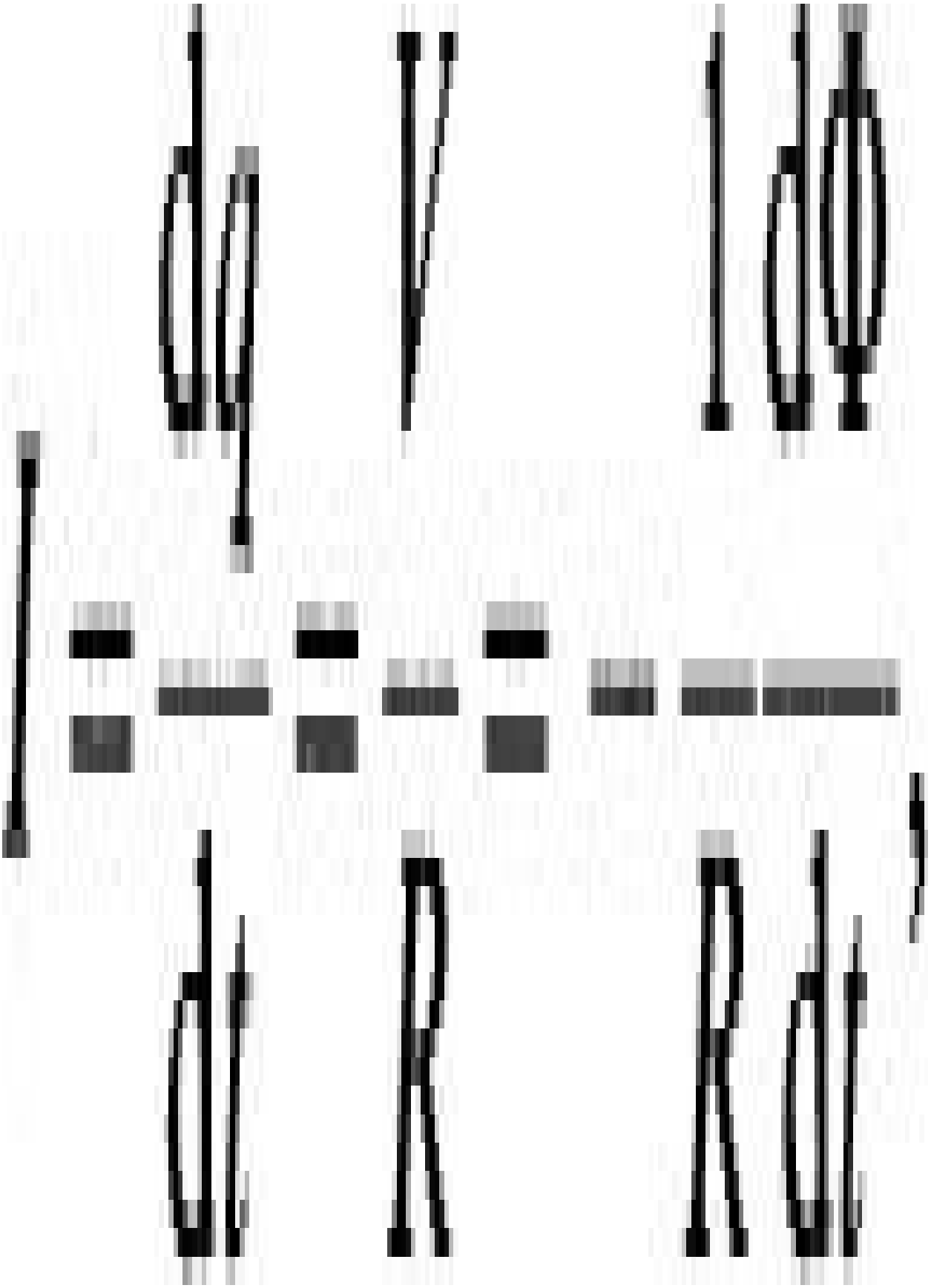
In this section we will discuss briefly some simple techniques for field measurement, and conclude by considering one method for the measurement of magnetic susceptibility.

6.4.1 The measurement of magnetic fields

Magnetic fields can be measured by determining the e.m.f.s induced when a coil placed in the field is removed.

While the coil is being removed from the field there is an e.m.f. V induced in the

circuit containing the coil. If the total resistance of the closed circuit which includes the coil is R, the induced current I is

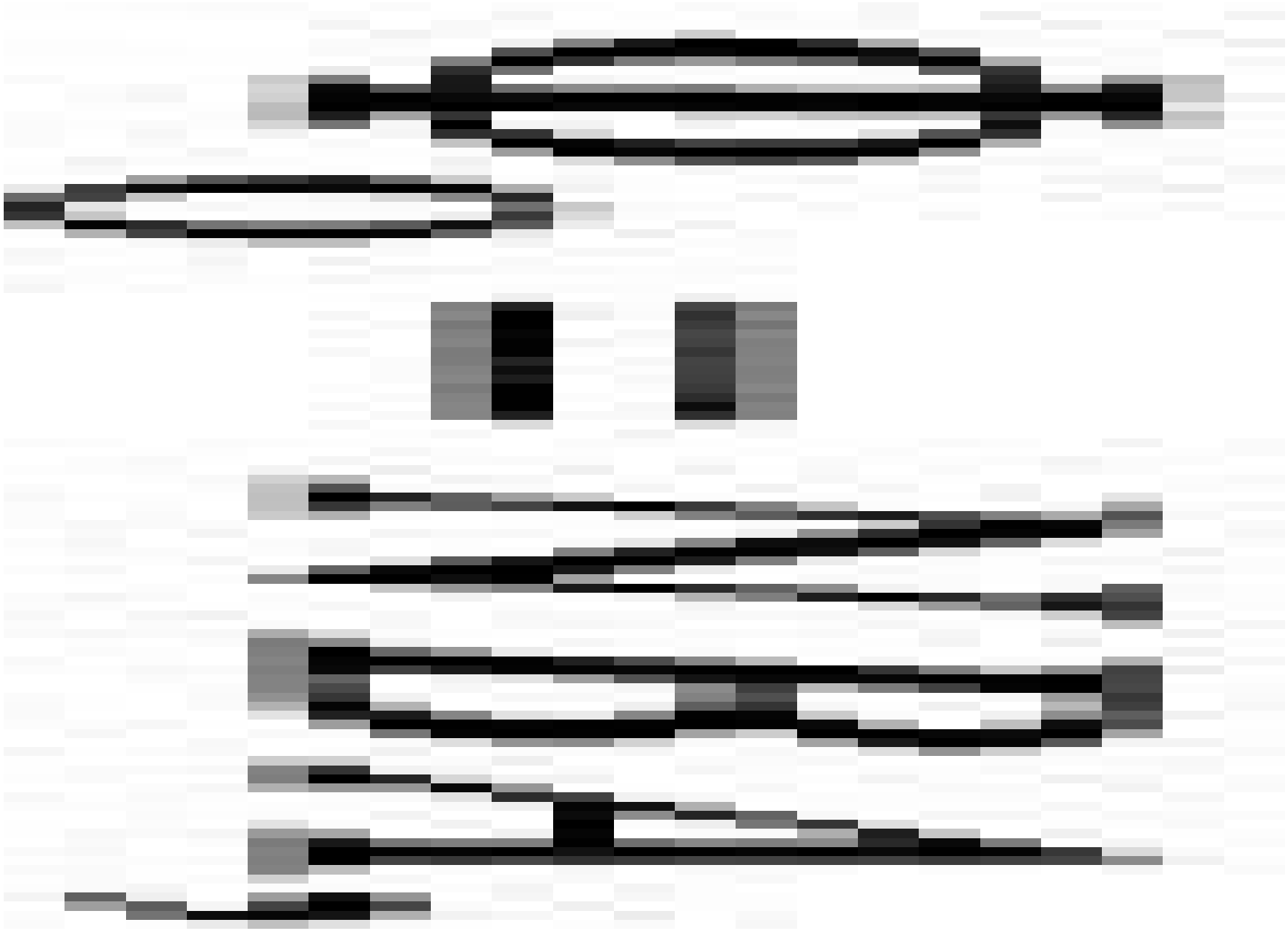


where Φ is the instantaneous total flux through the coil. The total charge that flows around the coil circuit during the time it takes to move the coil to a place where there is no field is

$$q = -\frac{1}{R}\int_0^{\infty} \frac{d\Phi}{dt}\,dt$$

$$= \frac{\Phi_0}{R},$$

where $\Phi 0$ is the original total flux through the coil. If the coil is small, so that the field B is uniform over the dimensions of the coil, and if the coil is perpendicular to the field, the magnetic flux is given by



where N is the number of turns on the coil, each of area A. Hence

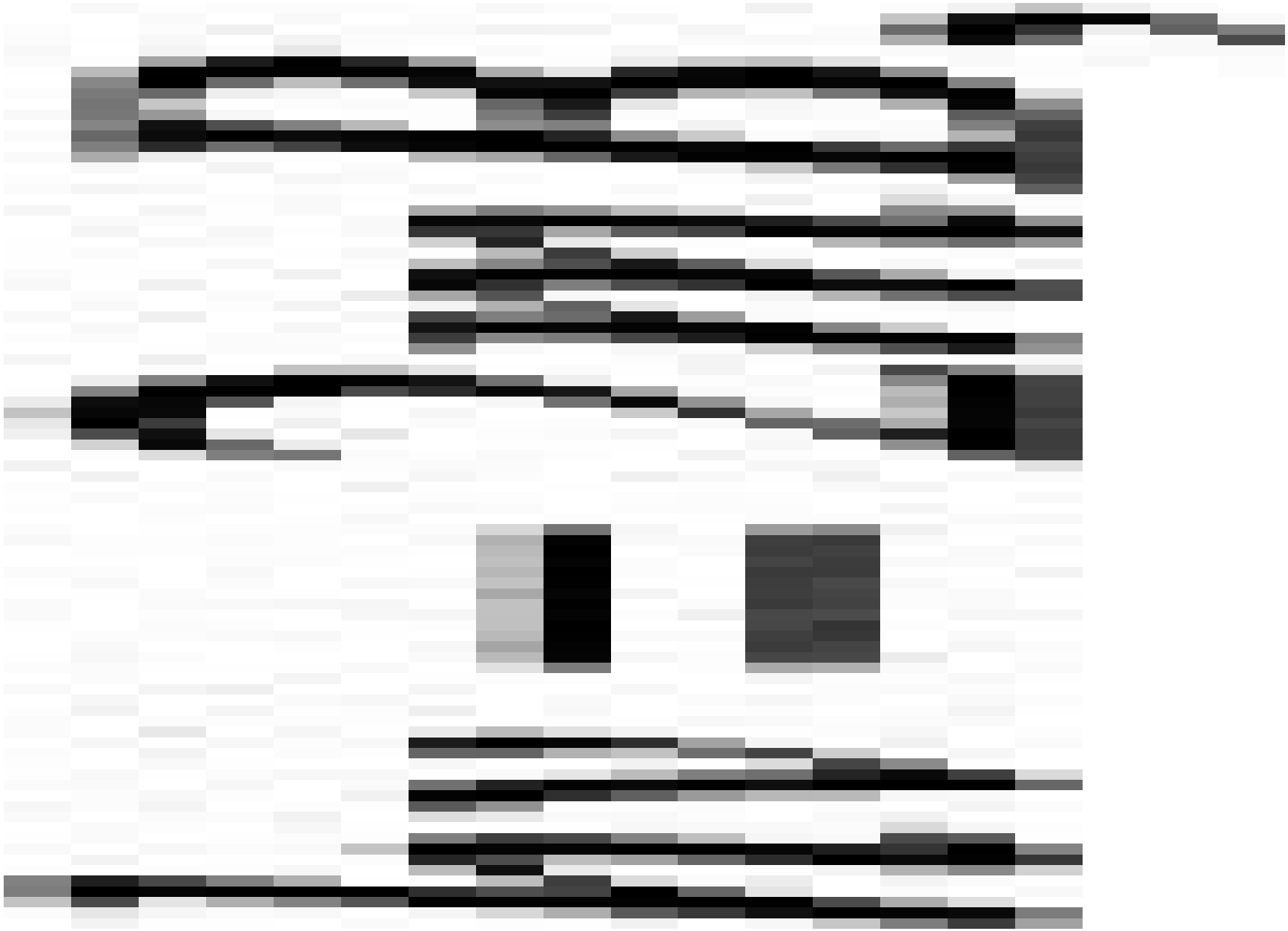
(6.41)

If the charge q is measured, and N and A known, the field B can be determined. One way of measuring the charge q is to use a suspended coil instrument. If the charge can be considered to have passed through the instrument before it has begun to rotate, it can be shown that the initial throw of the coil is proportional to the charge q.

Any physical effect which depends upon magnetic fields can be used in principle to measure them. Only some effects have the desired linear dependence on field strength and lend themselves to simple measurement techniques. One such, the Hall effect, has already been discussed in section 4.2.2. Most field measurements where high accuracy is not required are made using meters which measure the current resulting from the Hall effect. Hall effect meters are calibrated in magnetic fields accurately measured in other ways.

Another quantity which depends linearly on field is the potential energy of a magnetic dipole placed in the field. The proton has an intrinsic angular momentum quantum number equal to one half, and possesses a magnetic dipole moment m equal to + 2.793 μN. (One nuclear magneton, μN, is equal to one Bohr magneton multiplied by the ratio of the mass of the electron to the mass of the proton. Its numerical value is $5.051 \times 10{-}27$ A m^2. The plus sign means that the magnetic dipole points in the same direction as the intrinsic spin.) There will thus be two energy states for a proton in a diamagnetic material in a magnetic field B corresponding to the two allowed projections of the magnetic moment vector in the direction of the field. The two energy states have an energy difference equal to 2mB. The energy in a radio-frequency field of frequency v may be considered to reside in discrete units, called photons, each of which have

energy hv, where h is Planck’s constant. If a radio-frequency field of frequency v, given by



is applied to the sample it will induce transitions both up and down between the two energy states. If there are more transitions up than down there will be a net absorption of energy from the oscillating field, which can be detected. The frequency at which this occurs, as already implied, is in the radio-frequency region and can be measured very accurately. Hence the field can be determined in terms of the known magnetic moment of the proton. This technique is known as nuclear magnetic resonance, or NMR. The reason why there are more transitions up than down is that random interactions between the protons and other microscopic constituents of the material maintain statistical equilibrium. This means that the number of protons occupying the lower level is very slightly greater than the number occupying the upper level, according to the Boltzmann factor. The number of transitions upwards, proportional to the number of protons in the lower state, is thus greater than the number down, proportional to the number of protons in the upper level. This of course implies some deexcitation of the upper levels by transitions that do not involve emission of photons. This is done via the same mechanism responsible for the statistical equilibrium. The difference in energy between the two levels in a field of 0.1 T is about $2 \times 10{-}8\text{eV}$. The corresponding frequency at which absorption will occur is about 4.2 MHz. This is a frequency which can easily be measured to an accuracy of one part in 10^5. NMR is the technique most often used to measure magnetic fields accurately.

6.4.2 The measurement of magnetic susceptibilities

Accurate ways of measuring the very small magnetic susceptibilities of paramagnetic or diamagnetic substances depend upon the measurement of the force on a sample in a non-uniform magnetic field.

Suppose a long rod of the substance, whose cross-sectional area is A, is suspended vertically between the poles of an electromagnet, as shown in Figure 6.11. When the current is passed through the coils the rod will be attracted into

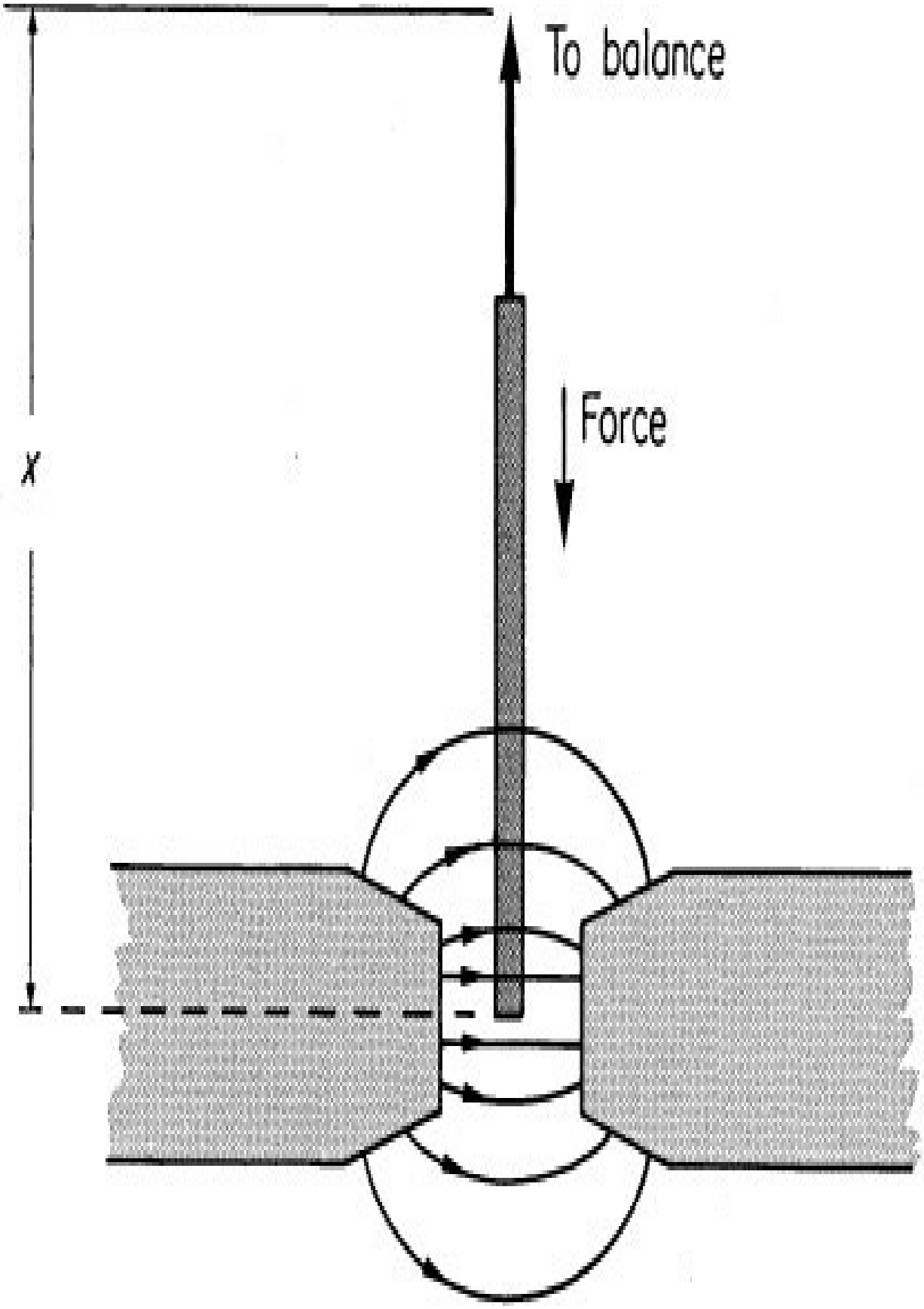
the field or repelled according to whether the material is paramagnetic or diamagnetic. This force in the vertical direction arises from the fringing field of the magnet where the field is non-uniform. However, we can find the magnitude of the force by calculating the change in potential energy of the sample when it is moved downwards through a small distance Δx. If the cross-sectional area of the rod is A, the net effect of such a displacement is to move an extra volume $A\Delta x$ into the region between the pole faces where the field B is uniform. Since the susceptibility χ_B of the rod is very small, it is a good approximation to assume that the field in the pole gap has the same value as in the absence of the rod, and that the rod acquires a uniform magnetization M equal to $\chi_B B/\mu_0$. The potential energy of a dipole m in a field B is given by Equation (4.26) as $-\mathbf{m}\cdot\mathbf{B}$. Hence the work done in building up the magnetization M is $= -\chi_B B^2/2\mu_0$ per unit volume. The magnetic potential energy UP of the rod is therefore changed by an amount $-\chi_B B^2 A\Delta x/2\mu_0$ when the rod moves through the distance Δx. The force F on the rod is given by

(6.42)

or

(6.43)

Figure 6.11. Weighing the force on a rod of paramagnetic material in a non-uniform magnetic field.

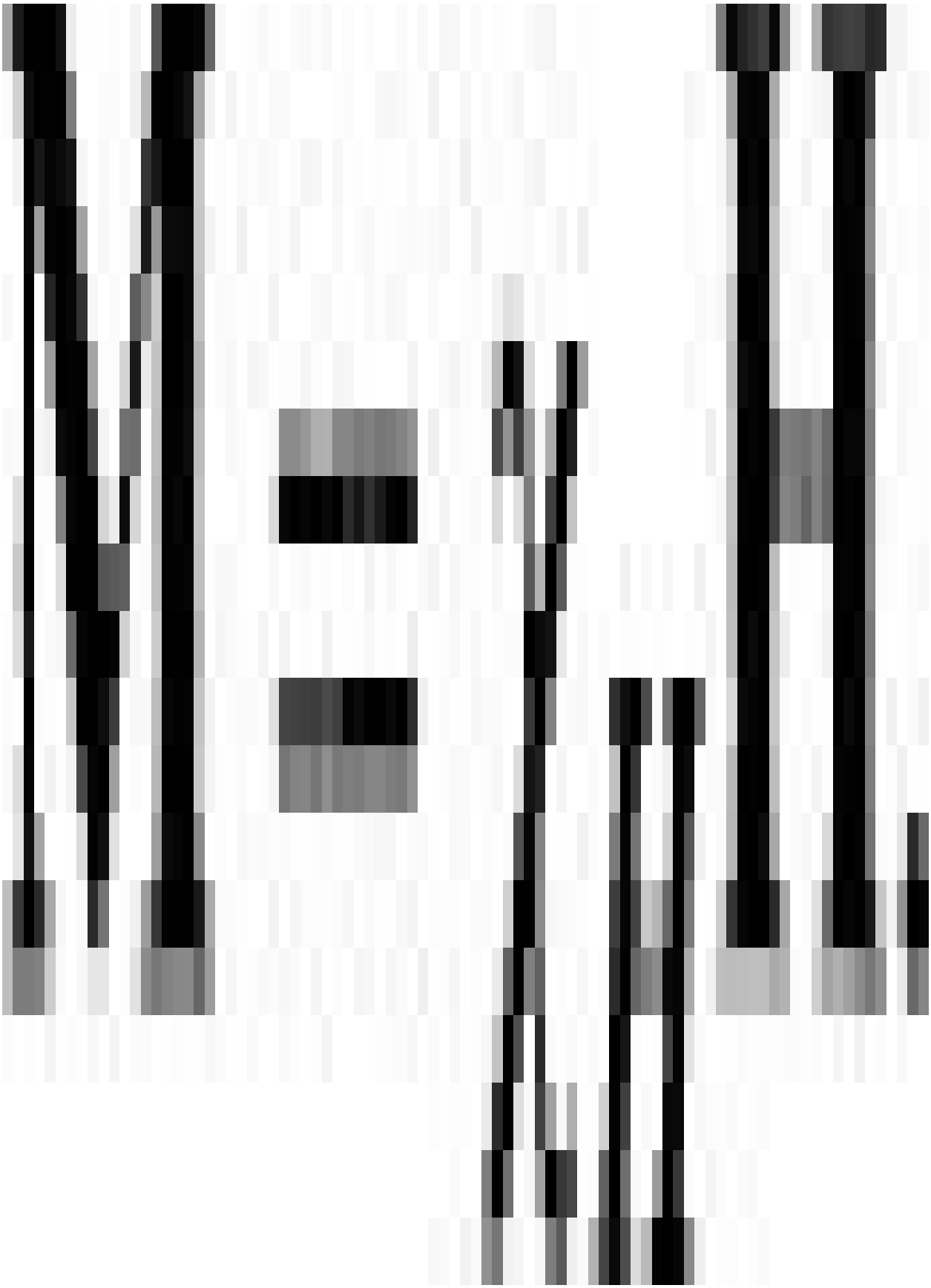

To balance
Force
X

This force is measured by weighing the rod using a sensitive balance, with and without current in the magnet coils, and the field B is measured in one of the ways described in the previous section.

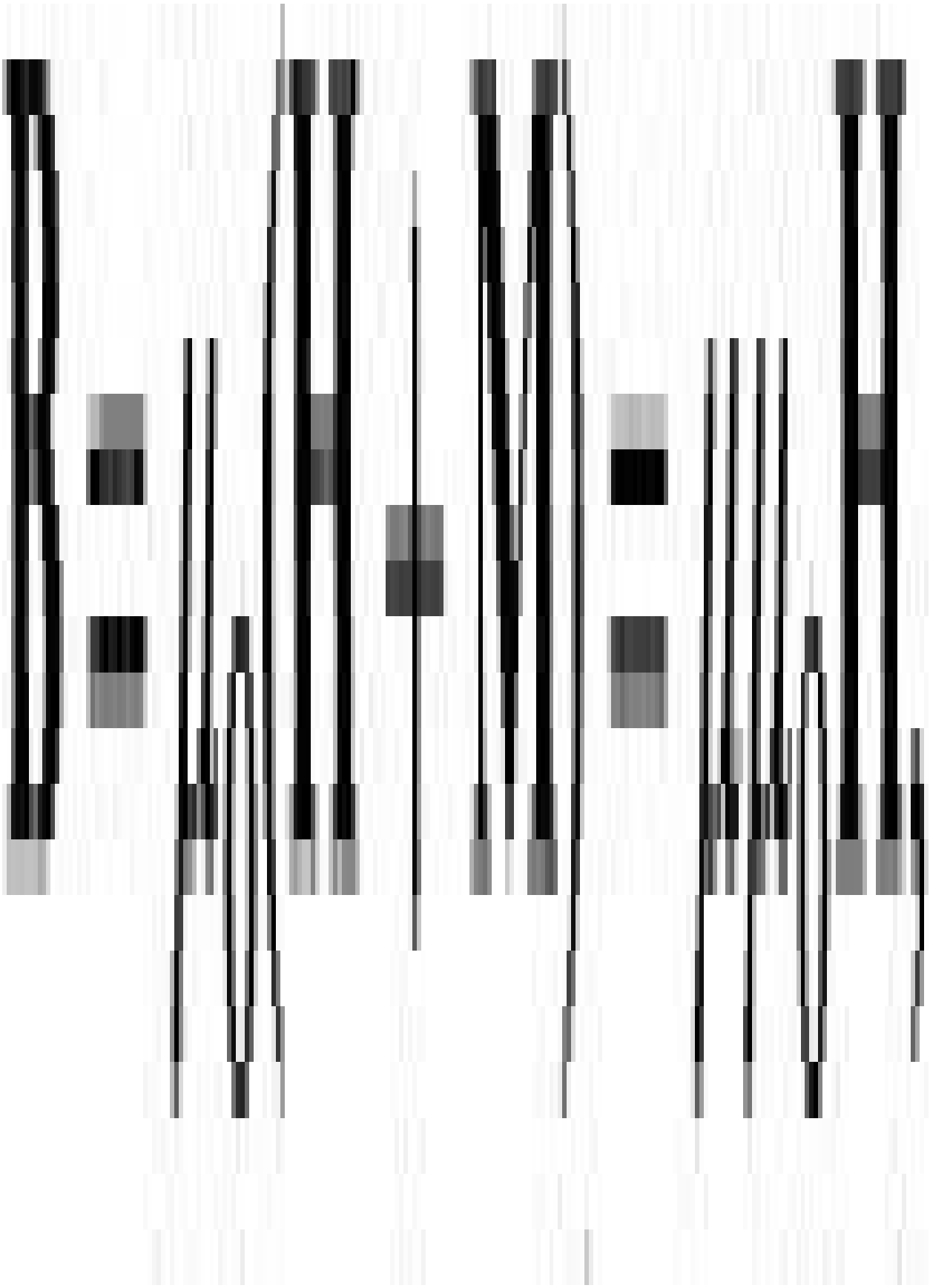
Aluminium is paramagnetic and has a susceptibility of $2.3 \times 10{-}5$. If an aluminium rod of cross-sectional area 1 cm^2 is suspended between the poles of an electromagnet which gives a field of 0.5 T, the force on the rod, given by Equation (6.43), is about $2.3 \times 10{-}4$N, i.e. about of the gravitational force on a mass of 1 g at the surface of the Earth.

As in the analogous electrical problem a local field correction is necessary in principle if one wishes to relate measurements of magnetic susceptibilities to atomic or molecular properties. The strength of the interaction between two typical molecular magnetic dipole moments is much smaller than that between two typical molecular electric dipole moments, and the approach adopted in section 2.2.3 to estimate the local electric field in gaseous and liquid dielectrics is valid for a wide sample of magnetic materials. The induced surface currents on the surface of a sphere cut out of a uniformly magnetized material give rise to a small reduction in the macroscopic field depending linearly on the susceptibility χB. However, since the susceptibility is so small the correction is negligible. For most practical purposes the magnetic fields inside volumes occupied by paramagnetic or diamagnetic materials are the same as they would have been if the materials were not there.

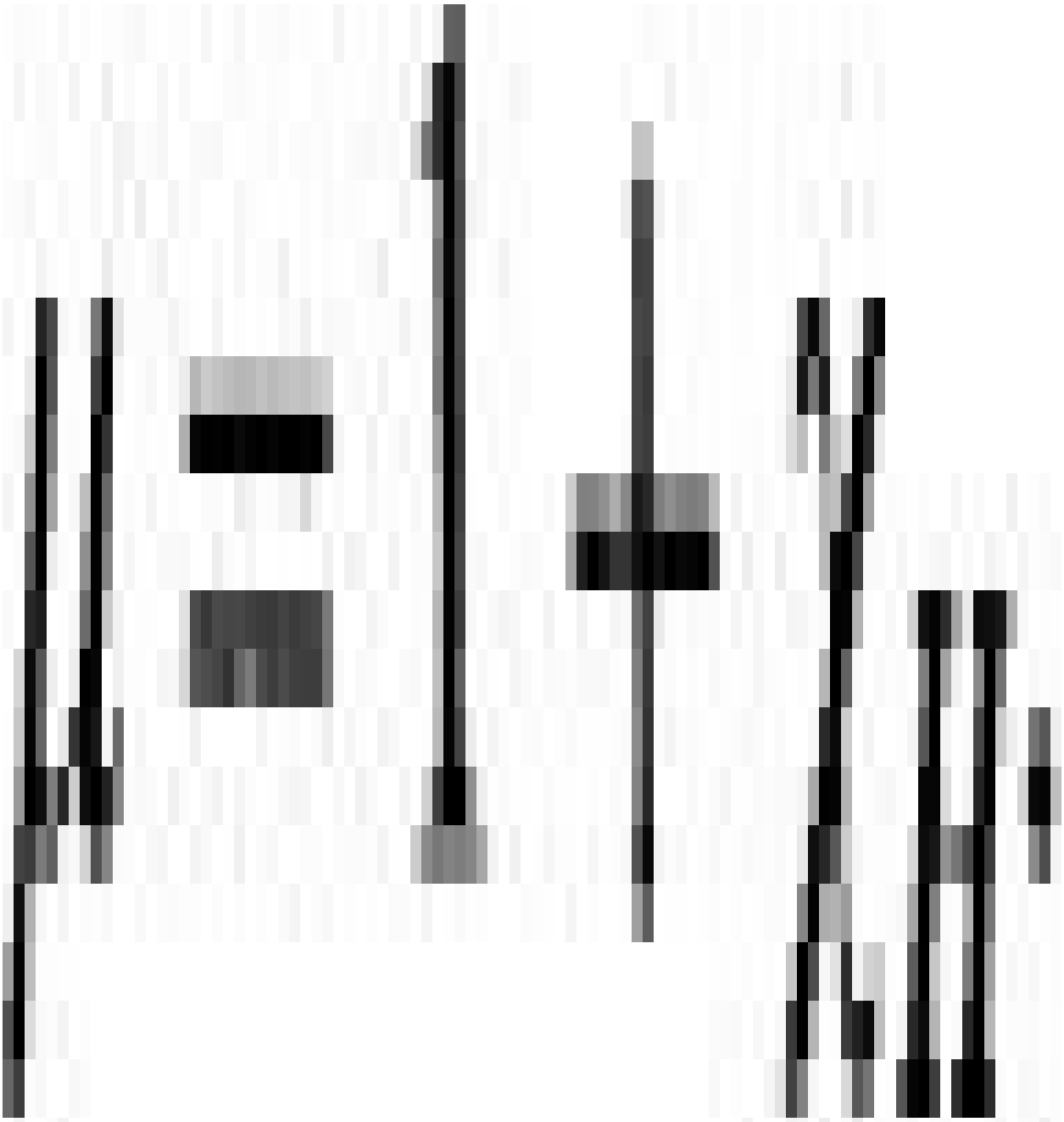
At the end of this section it is appropriate to point out that we have been using a different definition of magnetic susceptibility to the one usually used. The latter quantity, which we denote χH, is defined by relating the magnetization of a material to the magnetic intensity,



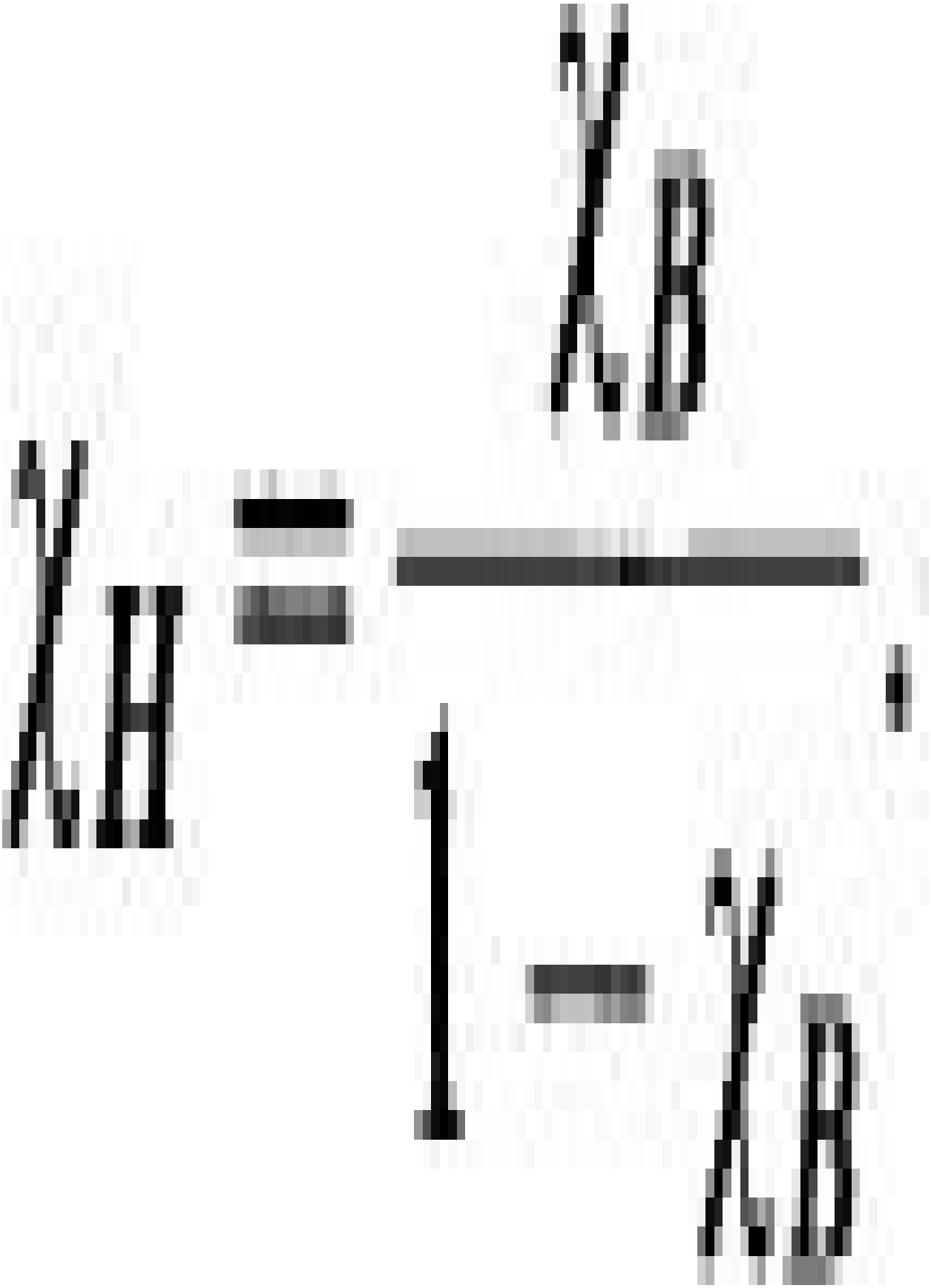
The magnetic field B is related to the fields M and H by



where μ is the relative permeability. Hence



From Equation (5.33) we may deduce the relationship between χ_H an χ_B;



For paramagnetic and diamagnetic materials the numerical difference between χ_H and χ_B is thus usually negligible. For ferromagnetic materials the difference is large; however for such materials it is not very useful to talk about magnetic susceptibility. It is more convenient to use the relative permeability μ, which itself is constant over only a very limited range of fields. The reason we choose to use χ_B instead of χ_H in this book is that it allows us to present a much more logical and satisfactory treatment of magnetic fields inside matter.

PROBLEMS 6

6.1 A rod of dielectric material is spun about its axis with angular velocity ω A uniform magnetic field B exists in a direction along the axis of the bar. Determine a charge distribution which produces the same electric field as does the rotating rod. The electric susceptibility of the material is χ_E.

6.2 A very small piece of magnetized material, whose magnetic dipole moment is M, pointing in the x direction, moves with speed v along the x-axis. A closed circular loop of thin wire, radius a, lies in the y–z plane at position x = 0. Determine the force opposing the motion of the particle. The speed v is much less than the speed of light, and the distance of the particle from the coil is much greater than the radius a. The resistance of the loop of wire is R, and it has negligible self-inductance.

6.3 When the armature of an electromagnetic relay is attracted, a closed iron circuit is formed, of cross-sectional area 4 cm^2, and length 20 cm. What current is required just to attract the armature if the initial gap is 1 cm, and a spring exerting a constant force of 10 newtons opposes the motion of the armature? The relay coil has 1000 turns and the permeability of the iron may be assumed to be constant at 500. To what value must the current be reduced before the armature opens again?

6.4 An iron toroid of permeability 1000 (which may be taken to be constant) has a cross-sectional area of 20 cm^2 and a mean radius of 50 cm. If a coil of 5000 turns is wound closely round the toroid, what would be the magnetic force between the faces of a gap 1 mm wide cut through the toroid with a current of 2 A passing through the coil?

What work must be done against the magnetic forces to increase the gap to 2 mm?

6.5 A coil of enamelled wire is wound onto a cylindrical insulating former. Describe and explain the effect on the self-inductance and the apparent resistance of the coil, when an alternating current is passed through it, of inserting into the solenoid

(i) a bar of iron

(ii) a bar of copper

(iii) a tube of copper

(iv) a tube of copper with a slit cut into it parallel to the axis of the tube.

6.6 If a steady e.m.f. is suddenly applied to a coil of self-inductance L1 show that, by the presence nearby of a closed second coil of self-inductance L2 which has a mutual inductance M with the first coil, the effective initial self-inductance of the first coil is diminished from L1 to $L1 - M^2/L2$.

6.7 A homopolar generator consists of a large metal flywheel rotating rapidly in a uniform magnetic field which is in a direction along the axis of the flywheel. The rotation causes a voltage difference between the axis and the rim of the flywheel. Homopolar generators are capable of producing very high currents for short periods of time of the order of a few seconds duration. Such high currents are used for example in experiments with plasmas (highly ionized gases) and for providing large magnetic fields in high energy nuclear physics machines. When the large currents are required through a coil, for example, the ends of the coil are connected to the rim and axis of the flywheel and the kinetic energy of the flywheel is converted into electrical energy. If a flywheel of mass 10^4 kg and radius 3 m is rotating at 3000 revolutions per minute in a field of 0.5 T, what would be the initial current delivered to a load of $10{-}3\Omega$? How long would it take for the flywheel to slow to half its speed?

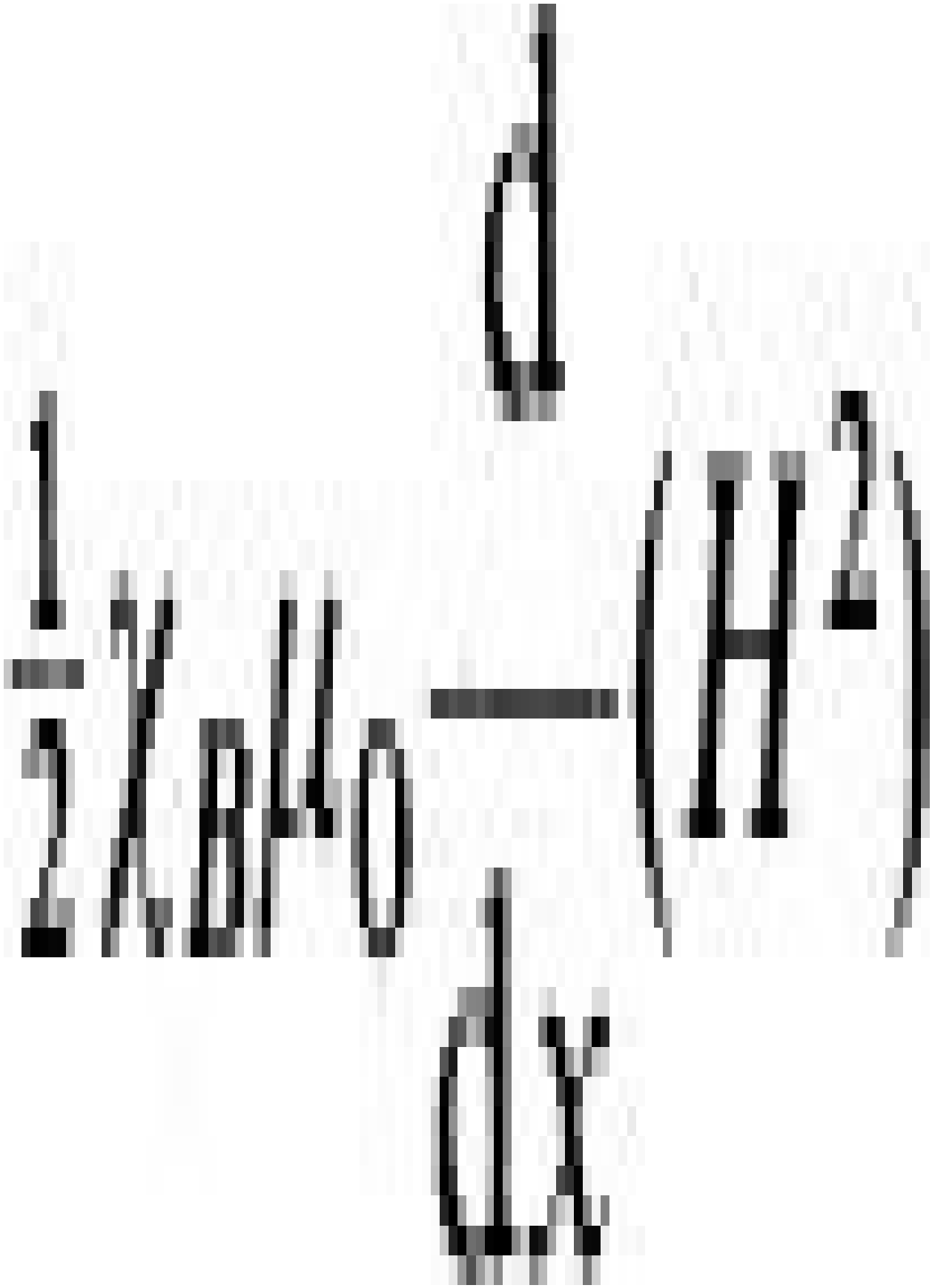
6.8 The coil of a moving coil instrument is connected in series with a circuit of resistance R and negligible self-inductance. Describe the motion of the coil if it is given a small displacement from its equilibrium position.

6.9 A coil of area A, resistance R and self-inductance L is rotated about a vertical axis in the plane of the coil with angular velocity ω The horizontal component of the Earth's magnetic field is B. Derive an equation for the current I in the coil and determine the mean torque required to rotate the coil.

6.10 Determine the total energy stored in the magnetic field of a long solenoid and hence show that the self-inductance is given by Equation (6.19).

6.11 Prove that the force per unit volume in the x-direction acting on a small

paramagnetic particle in an inhomogeneous magnetic field is given to a good approximation by



where χ_B is the magnetic susceptibility of the particle, and the particle is in air.

A rod of sulphur, 1.0 cm^2 cross-sectional area is placed in air in a non-uniform magnetic field, the direction of the field at each end of the specimen being at right angles to its axis. The strength of the field at one end a is 1.2 T and at the other end b 0.3 T. The force on the rod due to the field acts from a to b parallel to its axis and has magnitude 6.5×10^{-4}N. Find the magnetic susceptibility of sulphur.

6.12 A fluxmeter is a device sometimes used for measuring magnetic fields. It is a suspended coil instrument in which the suspension has virtually no restoring twist. The instrument is used in series with a search coil which is removed from or rotated in the field to be measured. The fluxmeter coil and the search coil form a continuous circuit of total resistance R. If the instrumental constant for the fluxmeter is k (see section 4.6.1), determine the deflection $\Delta\theta$ of the fluxmeter coil for a change of flux $\Delta\Phi$ through the search coil.

6.13 A steady magnetic field B is maintained inside a long solenoid of radius R. Determine the magnetic vector potential inside and outside the coil.

6.14 Determine the mutual inductance between a long straight wire and a rectangular loop with sides a and b. The loop is coplanar with the wire with its sides of length b parallel to the wire. The centre of the loop is a perpendicular distance x from the wire.

*See Chapter 14 for a discussion of how the electric and magnetic fields transform from one frame of reference to another frame in uniform relative motion.

*The flux through the orbit due to the motions of the other charges in the atom is to a good approximation constant and we neglect its effect in this analysis.

*The permeability of free space μ0 can be seen to have the dimensions of inductance divided by length. Its value is often quoted in units of henrys per metre ($\mu_0 = 4\pi \times 10^{-7}$ H m−1). The SI unit of magnetic flux is sometimes called the weber. One weber is equal to 1 T m^2. The unit of magnetic induction field B is sometimes quoted as webers m−2.

*The total work done by the batteries in fact includes another term due to Joule heating, which is an irreversible energy loss. This term does not concern us in this section, nor does it influence any of the conclusions reached.

*The situation discussed here is somewhat similar to that encountered in electrostatics when the charges on two conductors are changed whilst batteries maintain their potentials constant. The stored energy increases by an amount dU, while the batteries do work 2dU, corresponding to an amount of external work dU done by the system.

*We have ignored the fields due to the intrinsic magnetic dipole moments. The conclusion reached is still valid.

*All iron has hysteresis. In section 5.4.1 we implied that the relation between the fields B and H for soft iron was single valued. This is not a bad approximation for electromagnets, but when discussing energy losses in circuits which include iron, as now, we have to be more exact.

CHAPTER 7

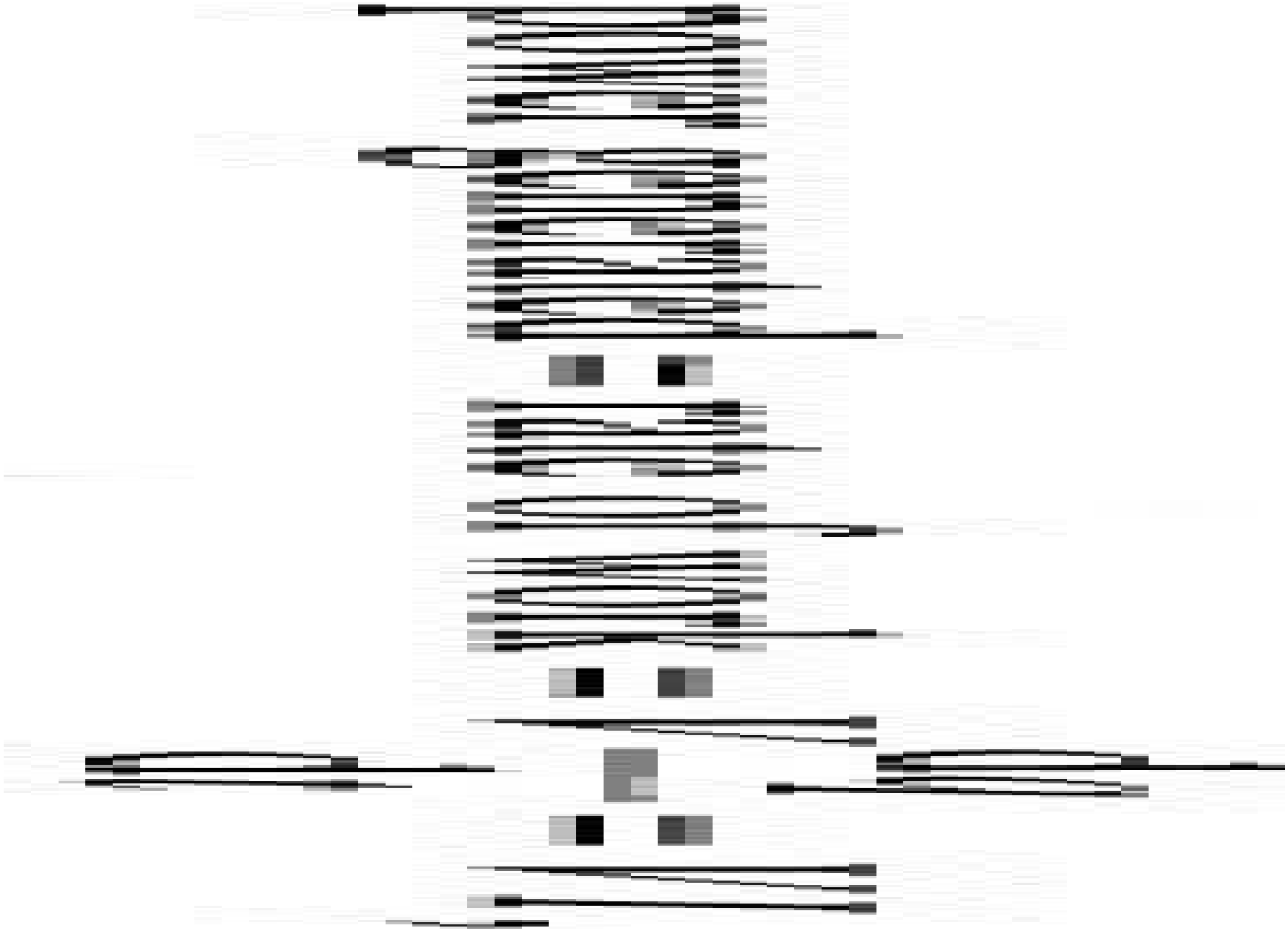
Alternating currents and transients

In the last chapter it was shown how an induced e.m.f. occurs whenever the magnetic flux through a circuit is changing. By introducing the concept of inductance, we were able to use Faraday's law to express induced e.m.f.s in terms of the rate of change of currents. We must now look at the opposite problem, and investigate the response of the currents in a circuit to known changes in the e.m.f.s driving them. This chapter is principally concerned with circuits in which both the currents and the e.m.f.s are varying sinusoidally with time. Such circuits are referred to as alternating current circuits, or A.C. circuits for short. There are good reasons for concentrating on the sinusoidal form of time variation. In the first place, many commonly occurring sources of e.m.f. do in fact vary sinusoidally. More importantly, in a circuit containing inductors and capacitors, sinusoidally varying e.m.f.s lead to sinusoidally varying currents, although as we shall see, the e.m.f.s and currents may not be in step with one another. The useful property of having the same functional form for e.m.f.s and current is not shared by any other type of time variation. Finally, according to Fourier's theorem, all functions may be expressed as a superposition of sine functions. It follows that we can predict the response of a circuit to any time variation of e.m.f. once we know its behaviour for alternating currents of all frequencies.

7.1 ALTERNATING CURRENT GENERATORS

Energy is dissipated when an electric current flows through a material with non-zero resistivity, and a continuous source of power is therefore needed to maintain a current in a circuit, except in the special case of superconductors. Consider the energy dissipated when a source of e.m.f. V drives a current I through a resistor. An amount of work Vdq is required to move a charge dq through the potential difference V, and

(7.1)



since the current I is the rate at which charge passes any point in the circuit. Electrical energy is not always dissipated as heat in a resistor: it may for example be partly converted into light, or used to drive an electric motor. However, the argument used to derive Equation (7.1) remains correct no matter how the energy is expended. At any moment, the instantaneous power of a source of e.m.f. V generating a current I is always given by the product VI.

In the alternating current generator, or dynamo, electrical energy is derived from mechanical work. The simplest dynamo consists of a single square coil rotating about an axis at right angles to a uniform magnetic field B, as shown in Figure 7.1. If the coil rotates with angular frequency ω, and is aligned at time zero so that B is in the plane of the coil, then the flux through the coil at time t is

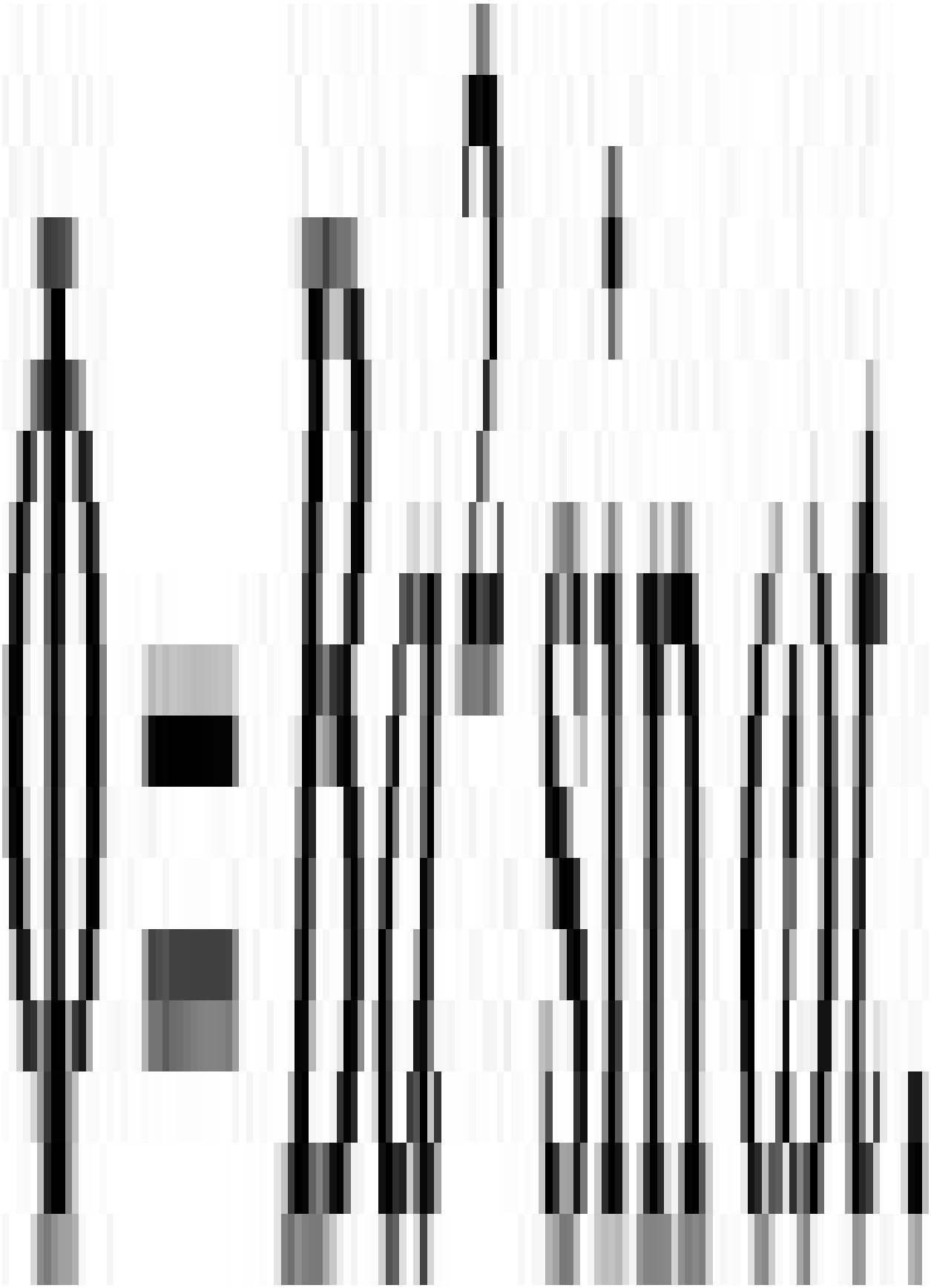
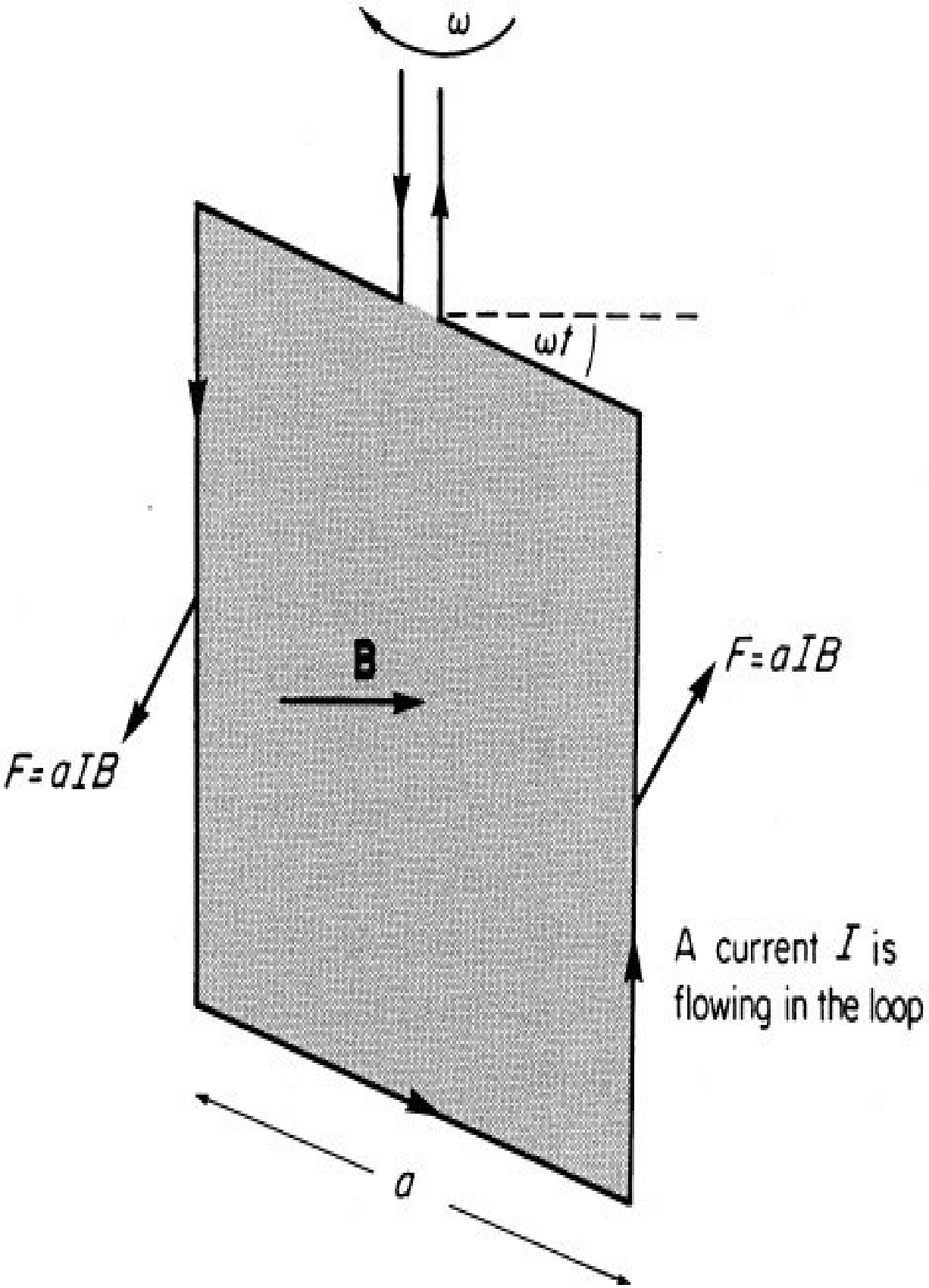
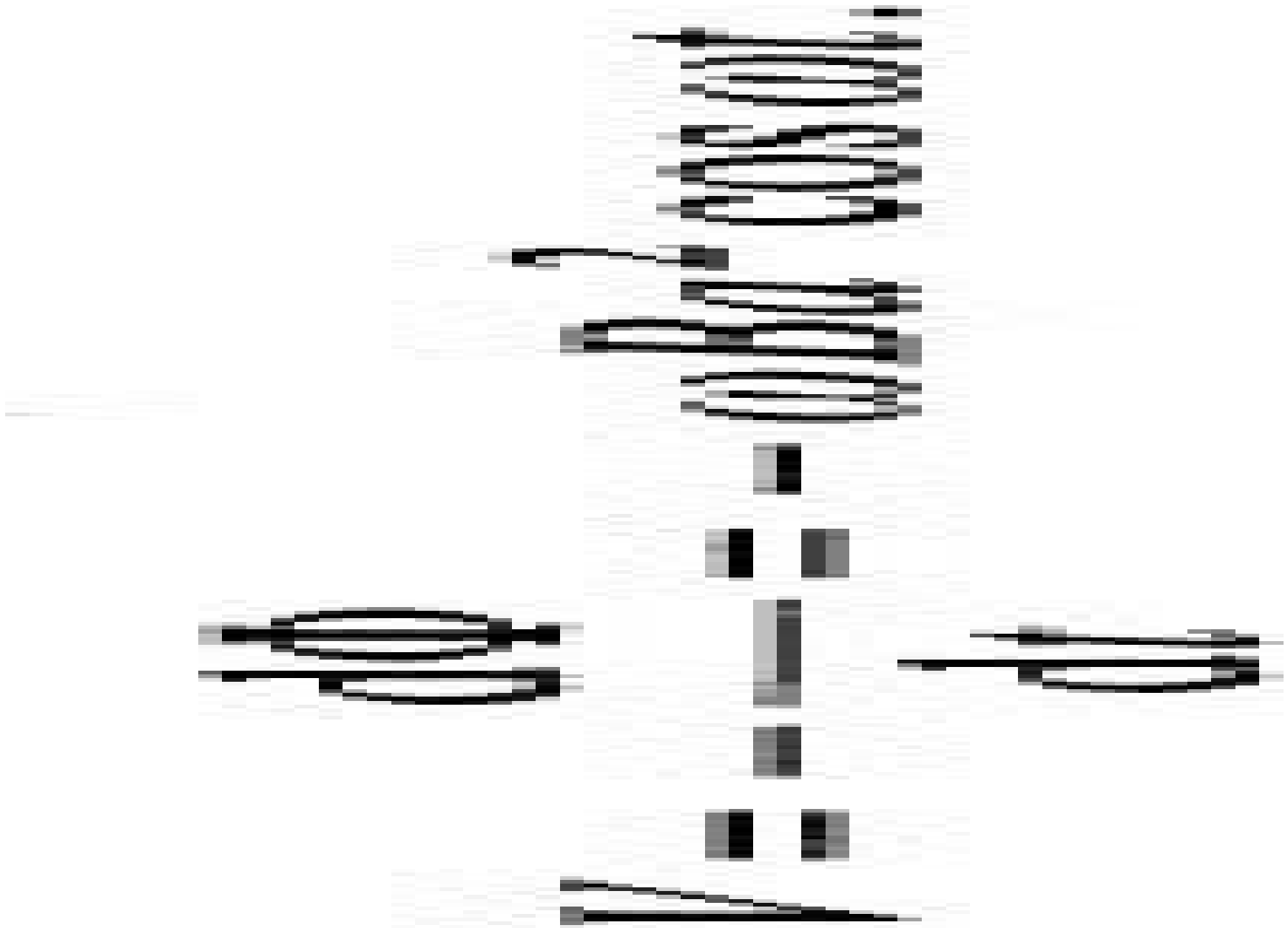


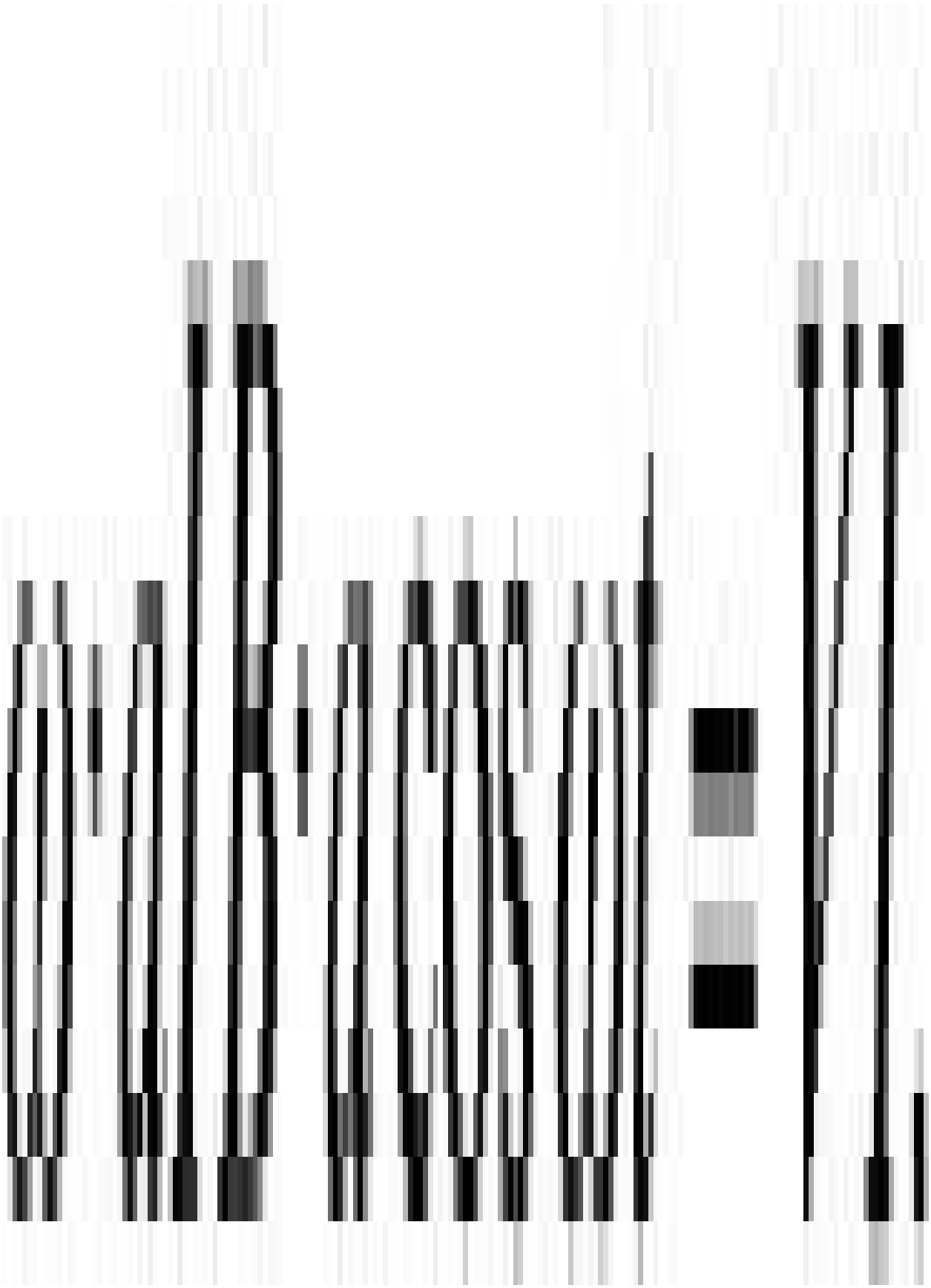
Figure 7.1. A simple dynamo made of a square loop of wire rotating in a magnetic field.


ω
ωt
B
F=aIB
F=aIB
A current I is flowing in the loop
a

Application of Faraday’s law tells us that the e.m.f. generated around the coil is



When the coil is open-circuited, so that no current flows, no power is developed by the generator, and no work is expended in rotating the coil. If the coil is made to form part of a closed circuit through which a current I is flowing, electrical power is generated. This power is derived from mechanical work, since there is now a torque opposing the motion of the coil. We can easily check that the work done against this torque is equal to the electrical energy generated by the coil. The current in the two vertical arms of the coil in Figure 7.1 is always perpendicular to B, and the Lorentz force on each vertical arm has a magnitude aIB. There is no resultant force or couple on the coil due to the forces on the other arms of the coil, and the net effect of the Lorentz force is a couple of magnitude ($aIB \cdot a \cos \omega t$) opposing the motion of the coil. The power required to maintain the rotation of the coil at a steady angular frequency ω is



The mechanical power consumed by the coil is thus exactly the same as the electrical power generated.

The large scale alternators used in power stations are similar in principle to the simple dynamo described here, although they contain many coils. The moving element is called the armature; it is a rotating electromagnet which supplies sinusoidally varying flux through stationary coils wound on the stator. Again mechanical work expended in driving the armature is converted into electrical energy, but not all of this energy is available to be used in external circuits. There is always some energy dissipated in the stator coils, which have a non-zero resistance. If the stator is delivering a current I, and the power lost in the coil is I^2r, then r is referred to as the internal resistance of the alternator. In practice r is usually small compared with the effective resistance of the external load, and alternators are therefore very efficient converters of mechanical energy to electrical energy. Nevertheless, when the alternator is driven, as it is in most power stations, by a heat engine—a steam or a gas turbine—the net efficiency of the station is a little less than 40%, because at practical operating temperatures the laws of thermodynamics limit the efficiency of the heat engine to about this value. In spite of this poor efficiency of conversion of fuel into electrical energy, only hydroelectric schemes and windmills, in which mechanical work is obtained directly from moving water or air, can compete economically with heat-engine driven alternators in the large-scale generation of electricity. Mechanical work is not the only form of energy which can be converted directly into electrical energy. For example, in a battery chemical energy is the source of power; some kinds of photoelectric devices maintain an e.m.f. when they are exposed to light; in a thermocouple heat energy drives a current without the intermediary of a dynamo. For large amounts of power these alternative sources of energy are not competitive. If you work out the cost of power in a flashlight you will find it to be at least a hundred times what is charged for the mains electricity supply.

7.2 AMPLITUDE, PHASE AND PERIOD

In the last section it was shown that a coil rotating with constant angular frequency ω radians/second in a uniform magnetic field generates an e.m.f. which varies sinusoidally with time. Let us choose the zero of time at a moment when an alternator generates its maximum e.m.f., which can then be written as

(7.2)

This voltage is drawn as a function of time in Figure 7.2. The maximum value V0 is called the amplitude of the e.m.f. and the argument ωt its phase. The cosine pattern repeats itself every time ωt increases by 2π, and the phase is usually restricted to the range 0 to 2π. When ωt reaches an integral multiple of 2π, the phase begins again at zero. The voltages at t1 and t2 in Figure 7.2, for example, are said to be at the same phase. The period between successive repetitions of the same phase is

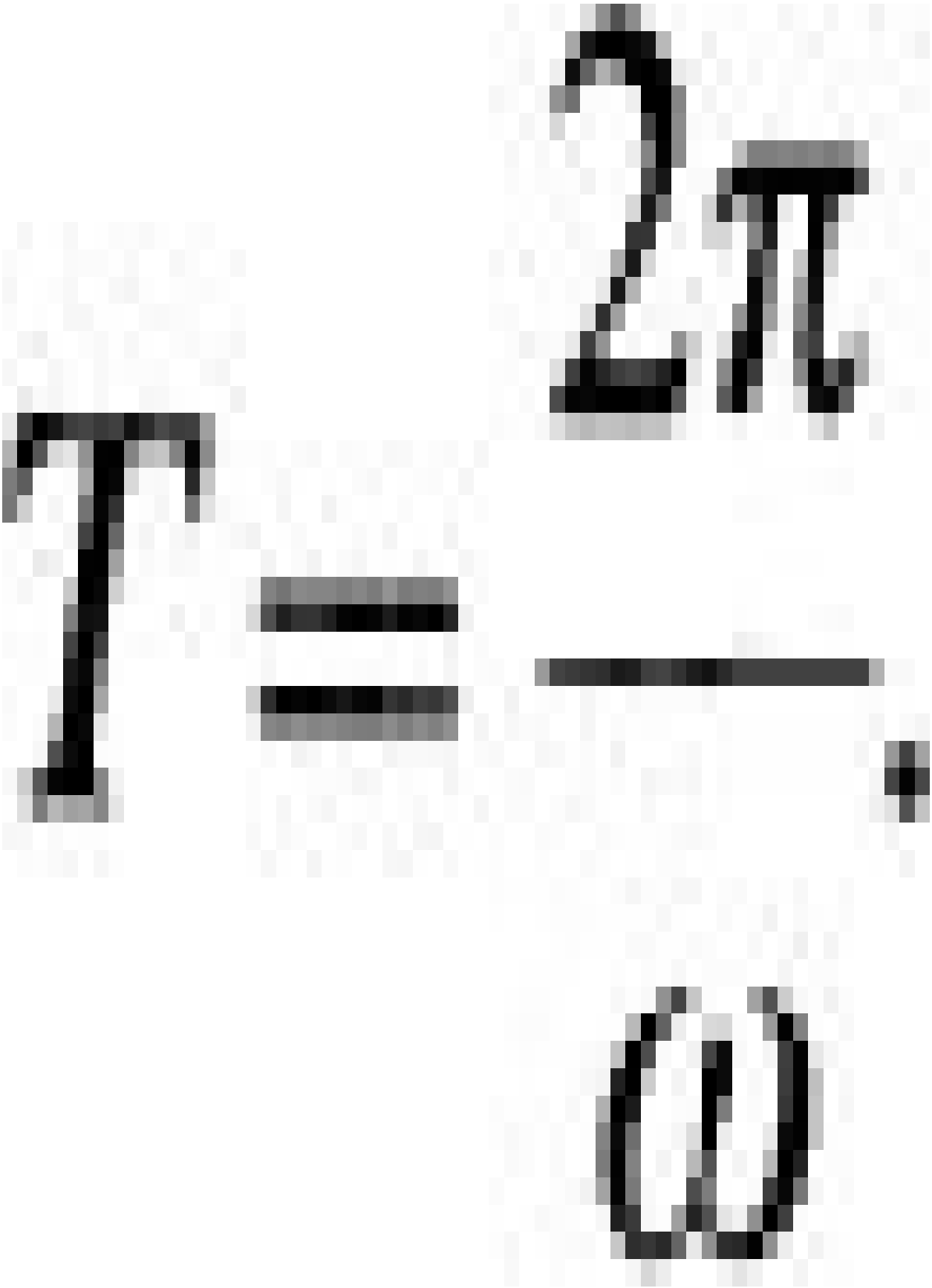
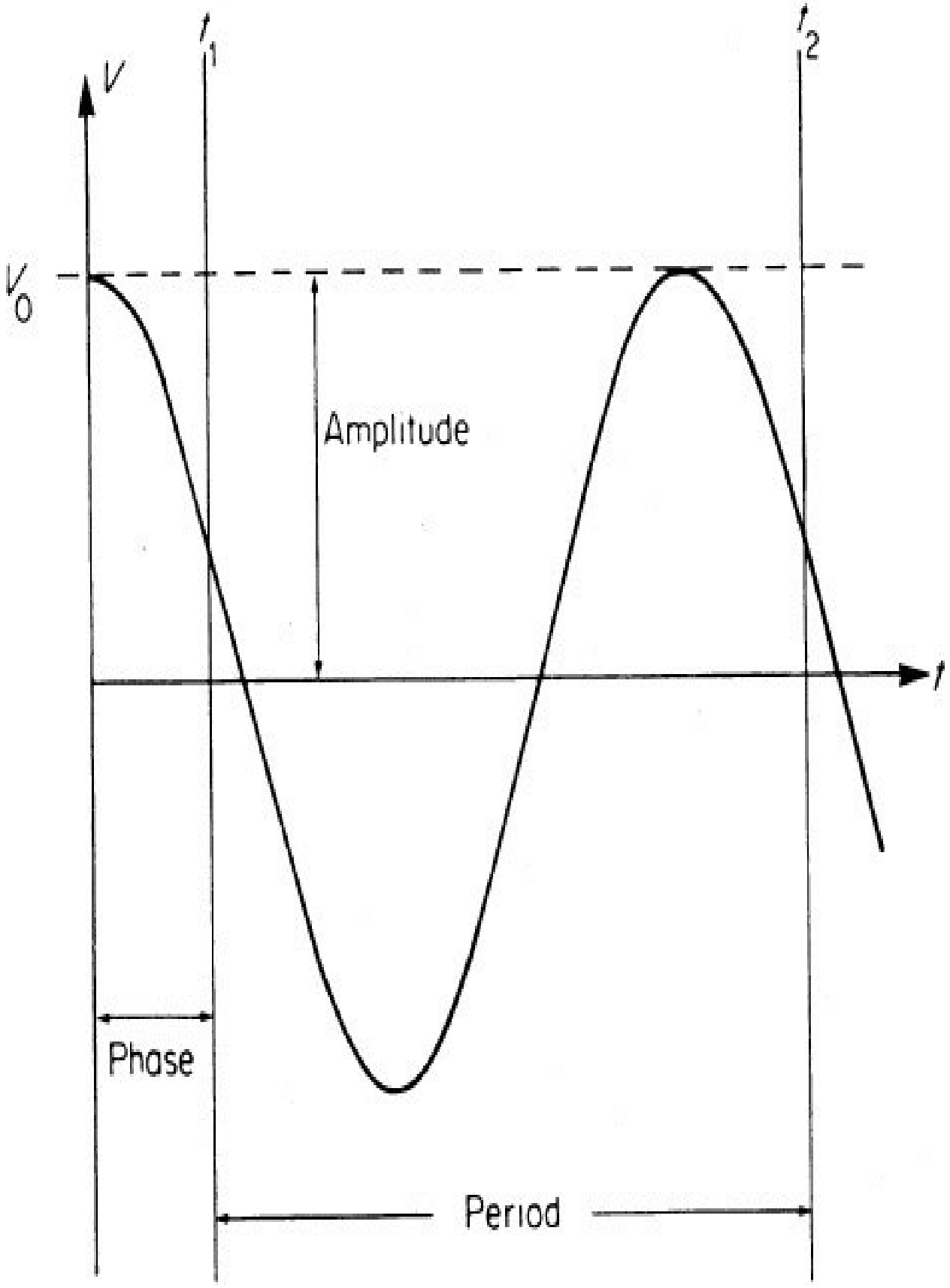
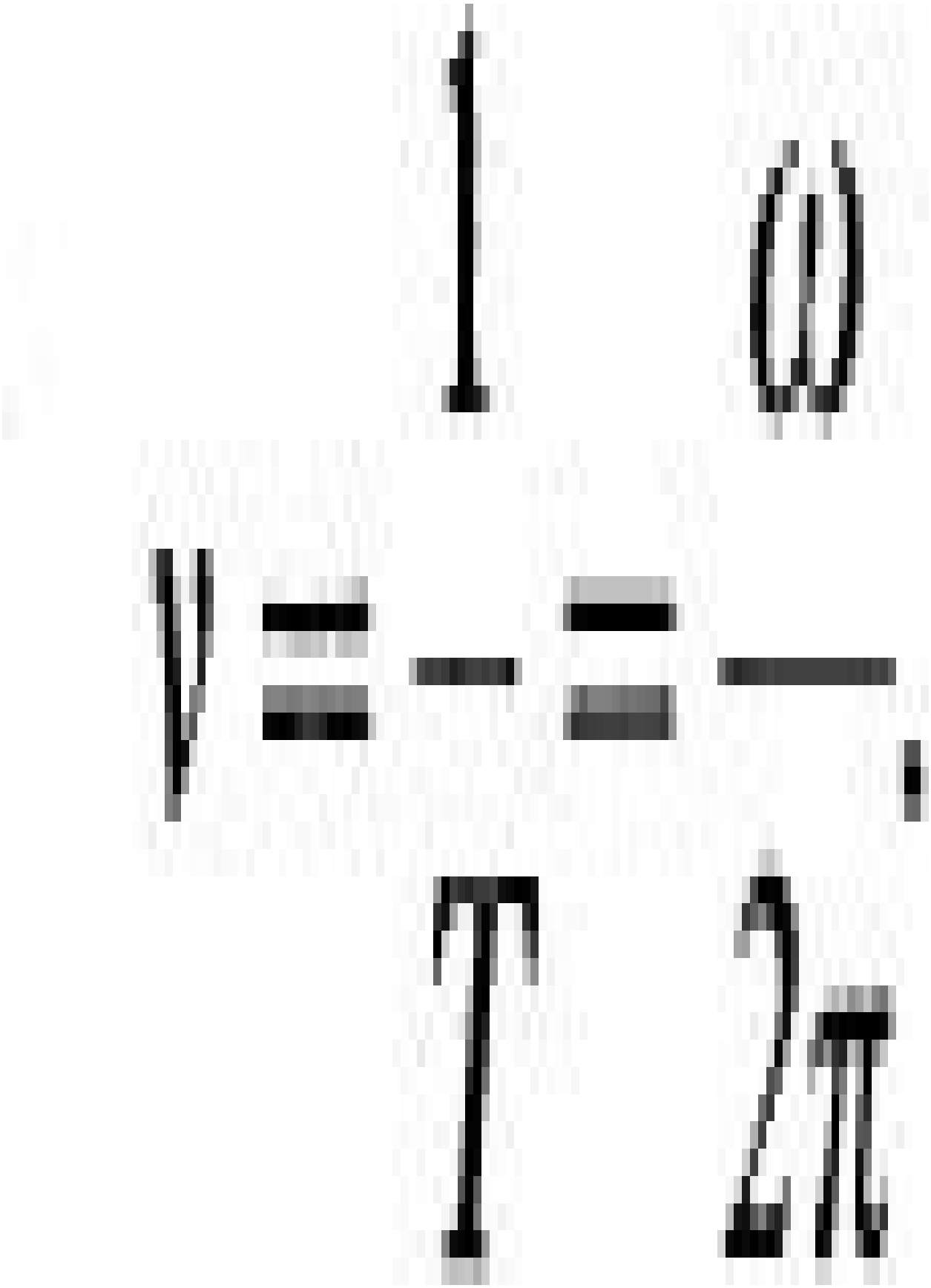


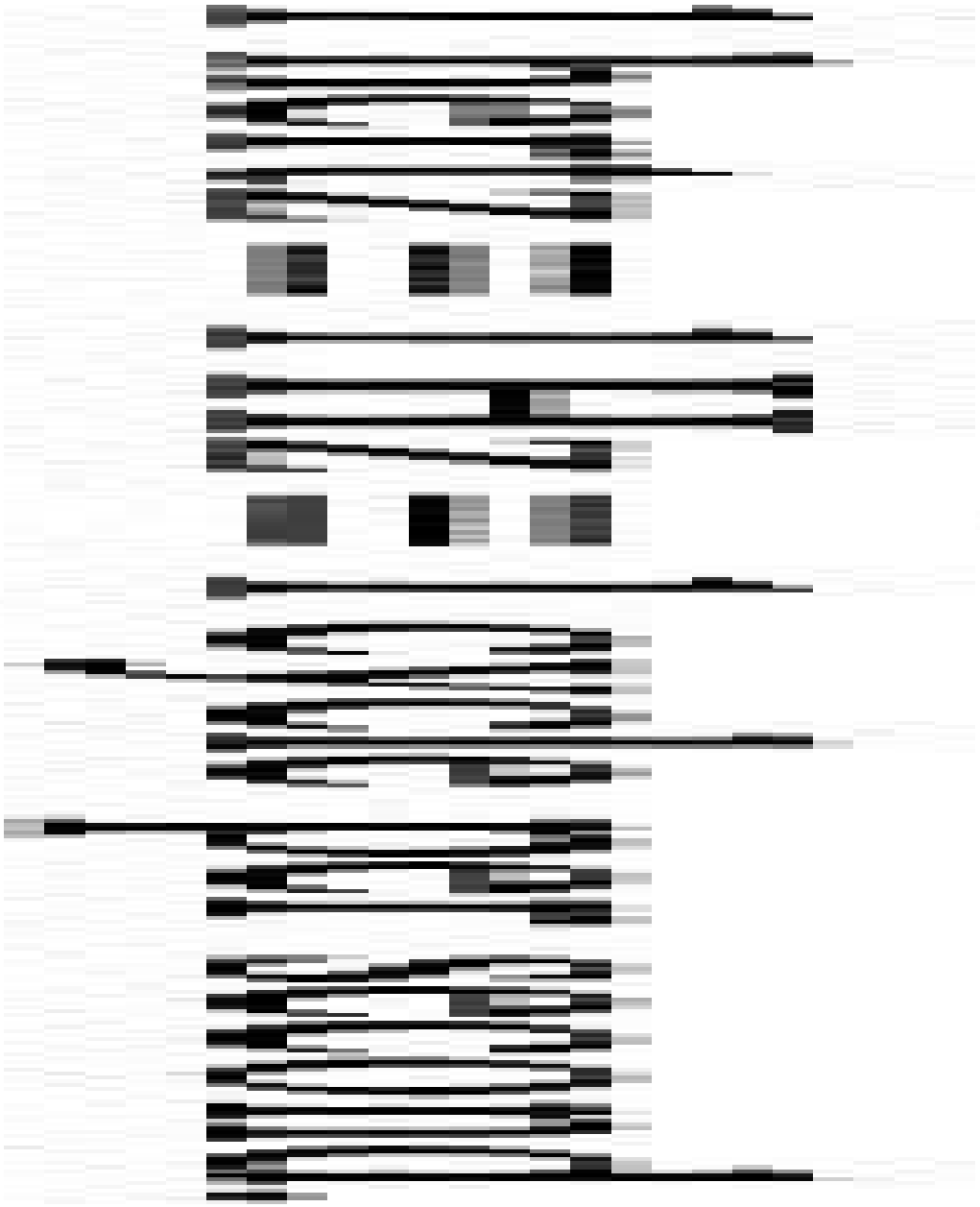
Figure 7.2. The alternating voltage waveform.


V
t_1
t_2
V_0
Amplitude
t
Phase
Period

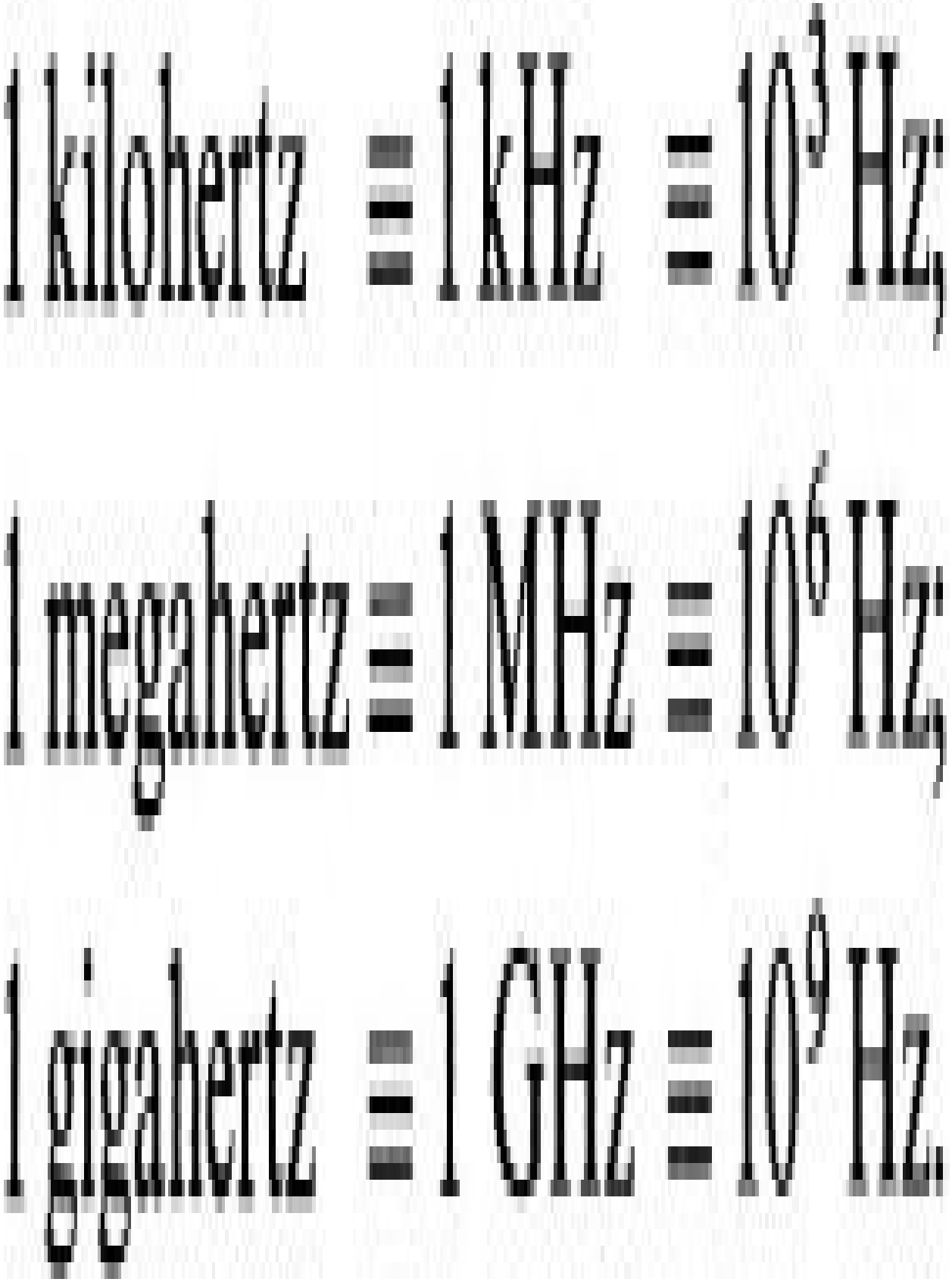
The sequence of events occurring in one complete period is called the cycle of the generator. The voltage pattern repeats at a frequency



There is a special unit of frequency called the hertz, after the discoverer of radio transmission:



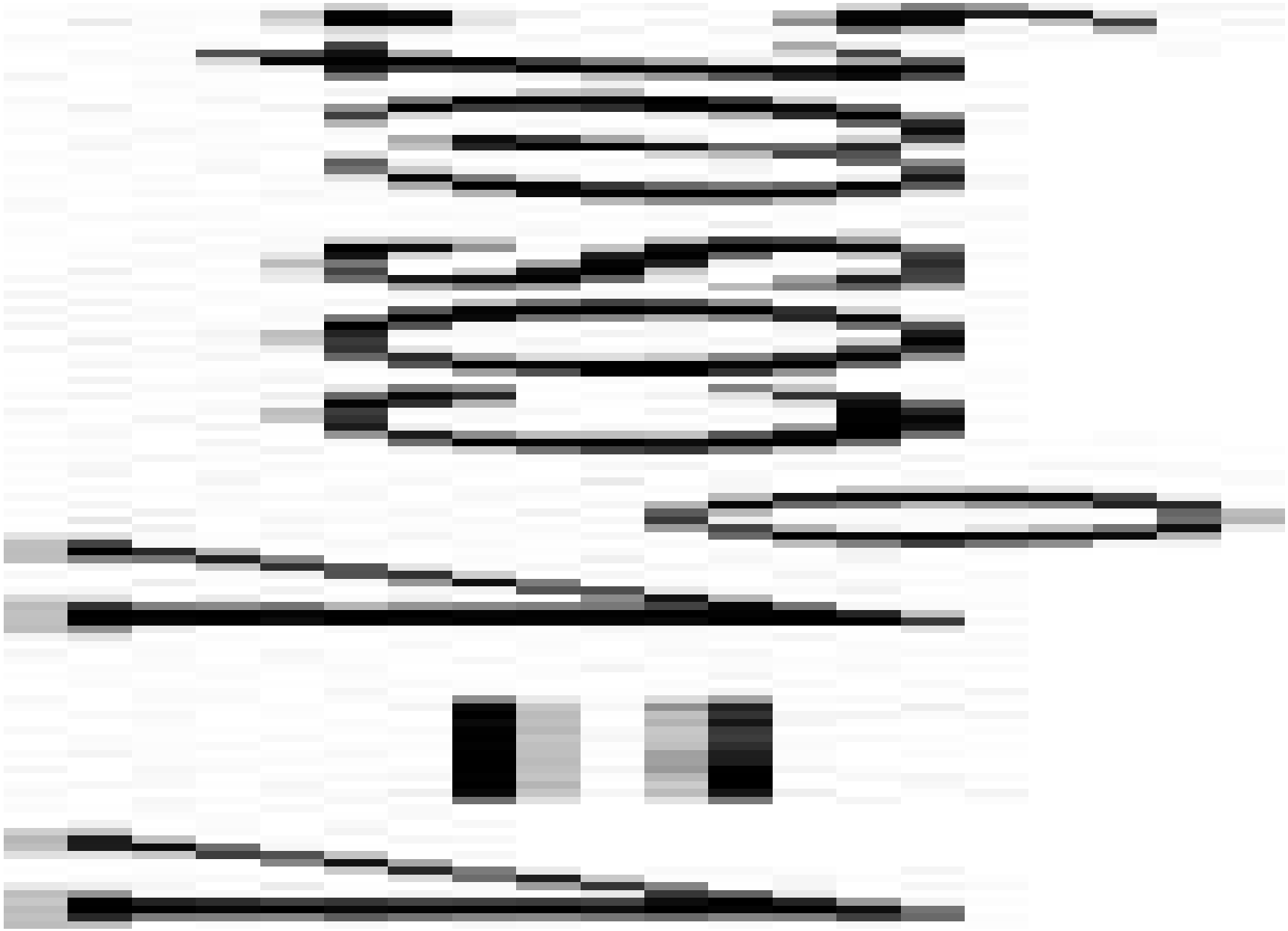
The standard units for multiples of the hertz are


1 kilohertz = 1 kHz = 10³ Hz
1 megahertz = 1 MHz = 10⁶ Hz
1 gigahertz = 1 GHz = 10⁹ Hz

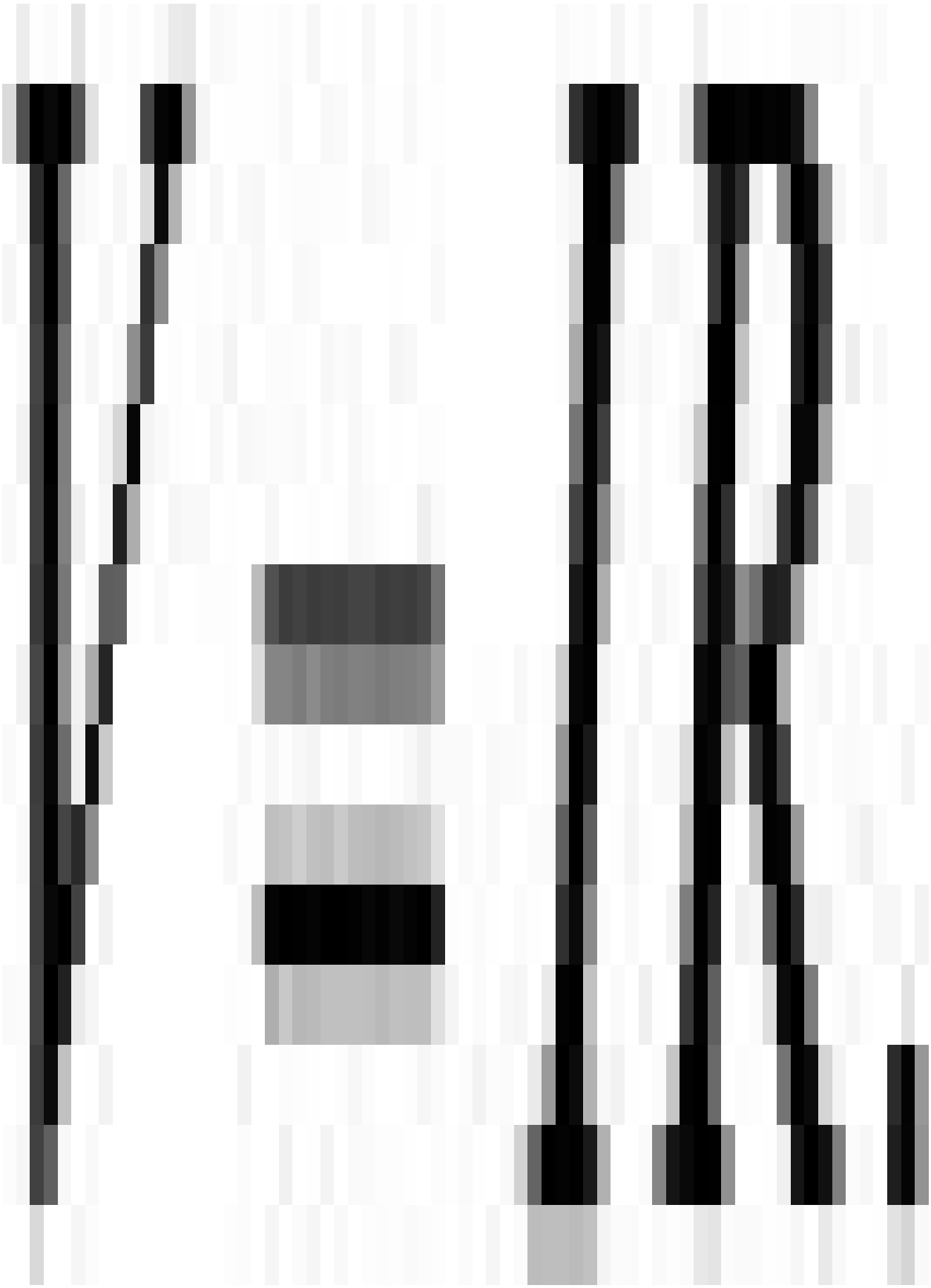
The frequency of the main electricity supply is 50 Hz in Europe, and 60 Hz in the United States. Alternating voltages of much higher frequency occur in the tuned circuits used in telecommunications, and it is important to know how the circuits respond to the application of a sinusoidally varying e.m.f. of any frequency. We shall begin the study of A.C. circuits by considering the currents driven by an A.C. generator through single circuit elements.

7.3 RESISTANCE, CAPACITANCE AND INDUCTANCE IN A.C. CIRCUITS

Figure 7.3(a) shows a circuit consisting of an A.C. generator (indicated by the symbol) driving a current I through a resistor. The generator delivers an oscillating e.m.f.



the resistor obeys Ohm’s law and has a resistance R, so that



Hence the current through the resistor is

(7.3)

The current is in phase with the applied voltage, and has an amplitude $I_0 = V_0/R$.

The next diagram, Figure 7.3(b), shows the same generator connected across a capacitor of capacitance C. Again the voltage is $V_0 \cos \omega t$, and it is now related to the charge q carried by the capacitor through the equation

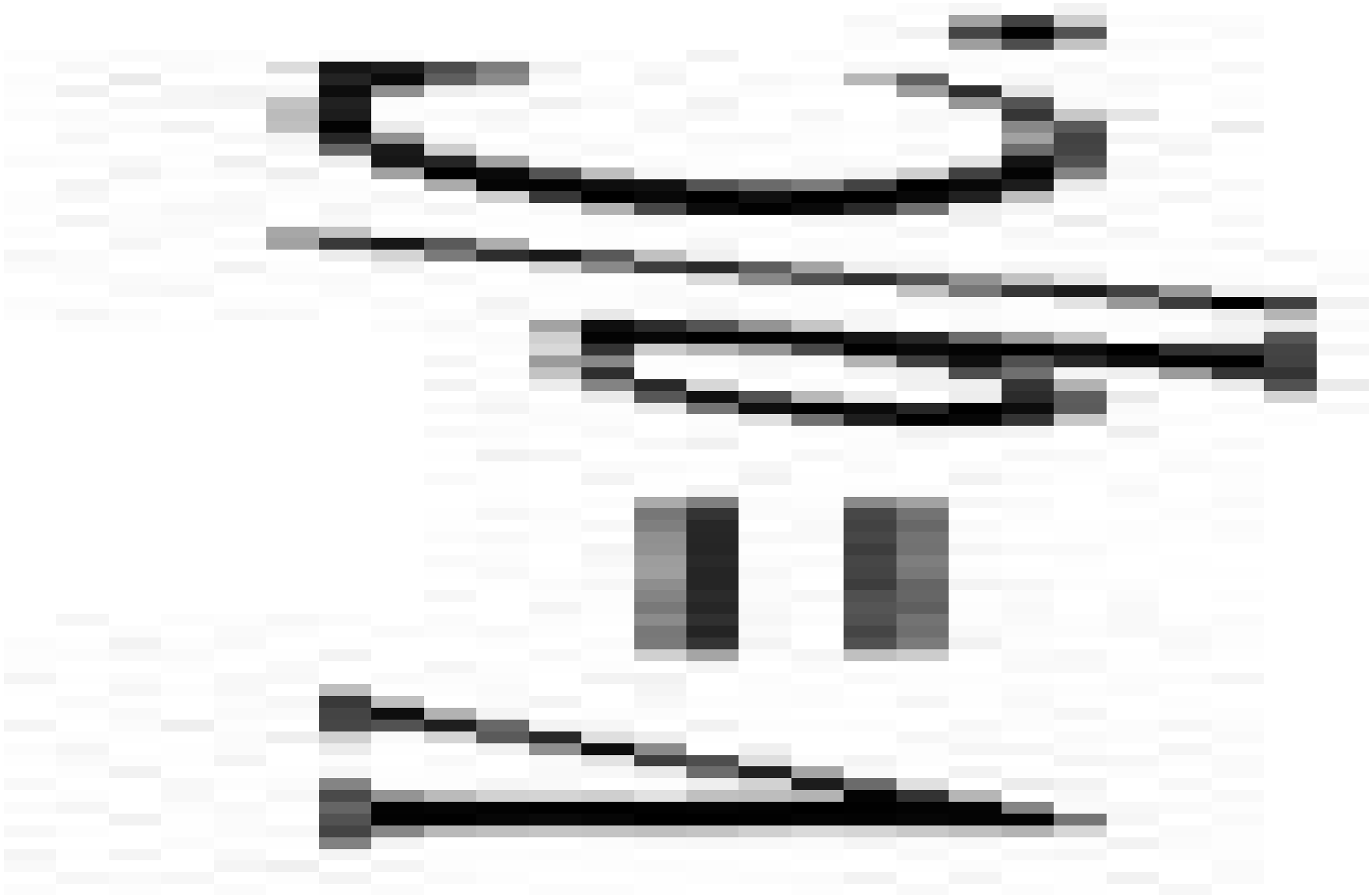
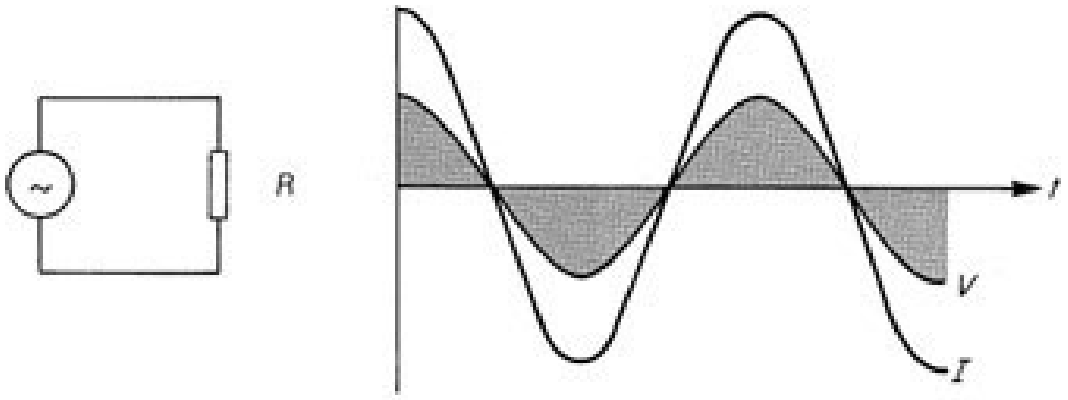
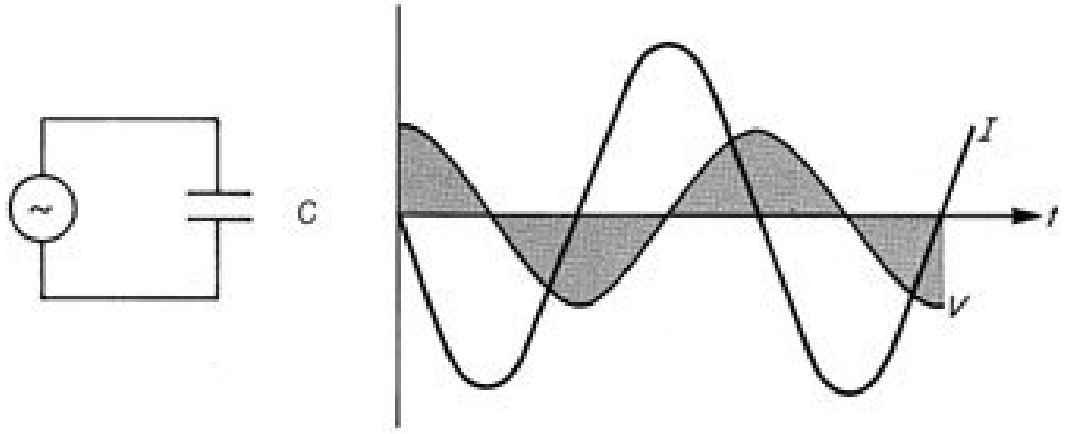


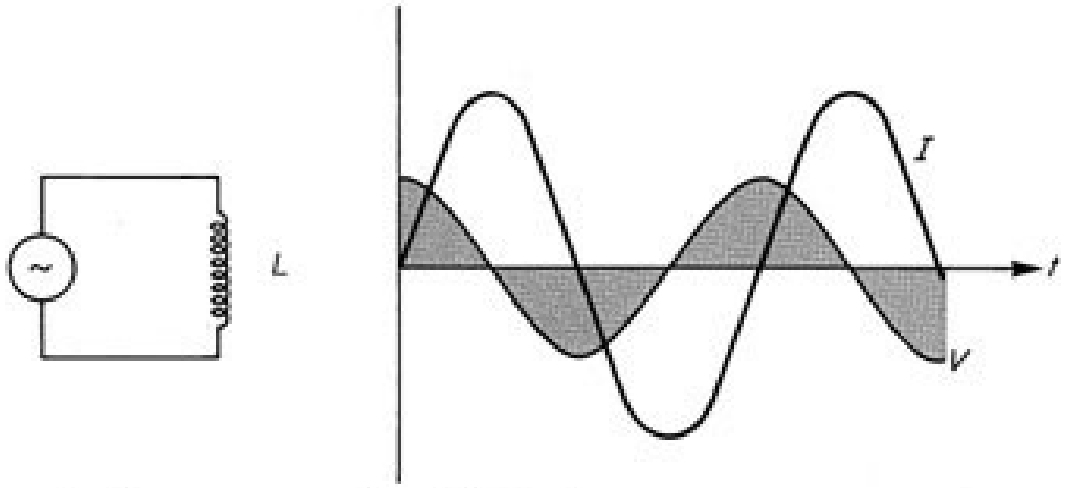
Figure 7.3. Phase relations between current and voltage.



(*a*) Resistance: current and voltage are in phase.

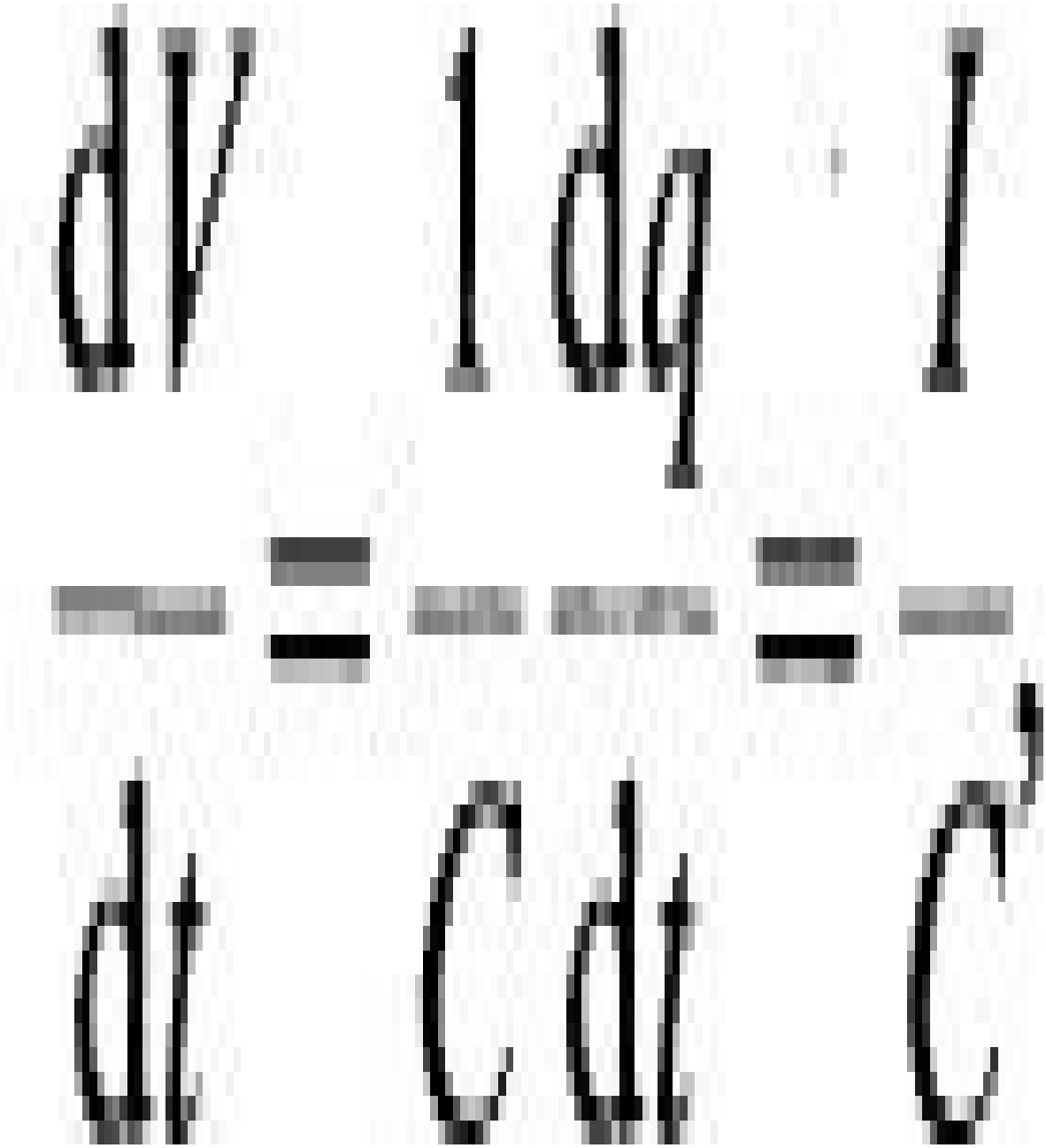


(*b*) Capacitance: current leads voltage.



(*c*) Inductance: current lags behind voltage.

Differentiating we find

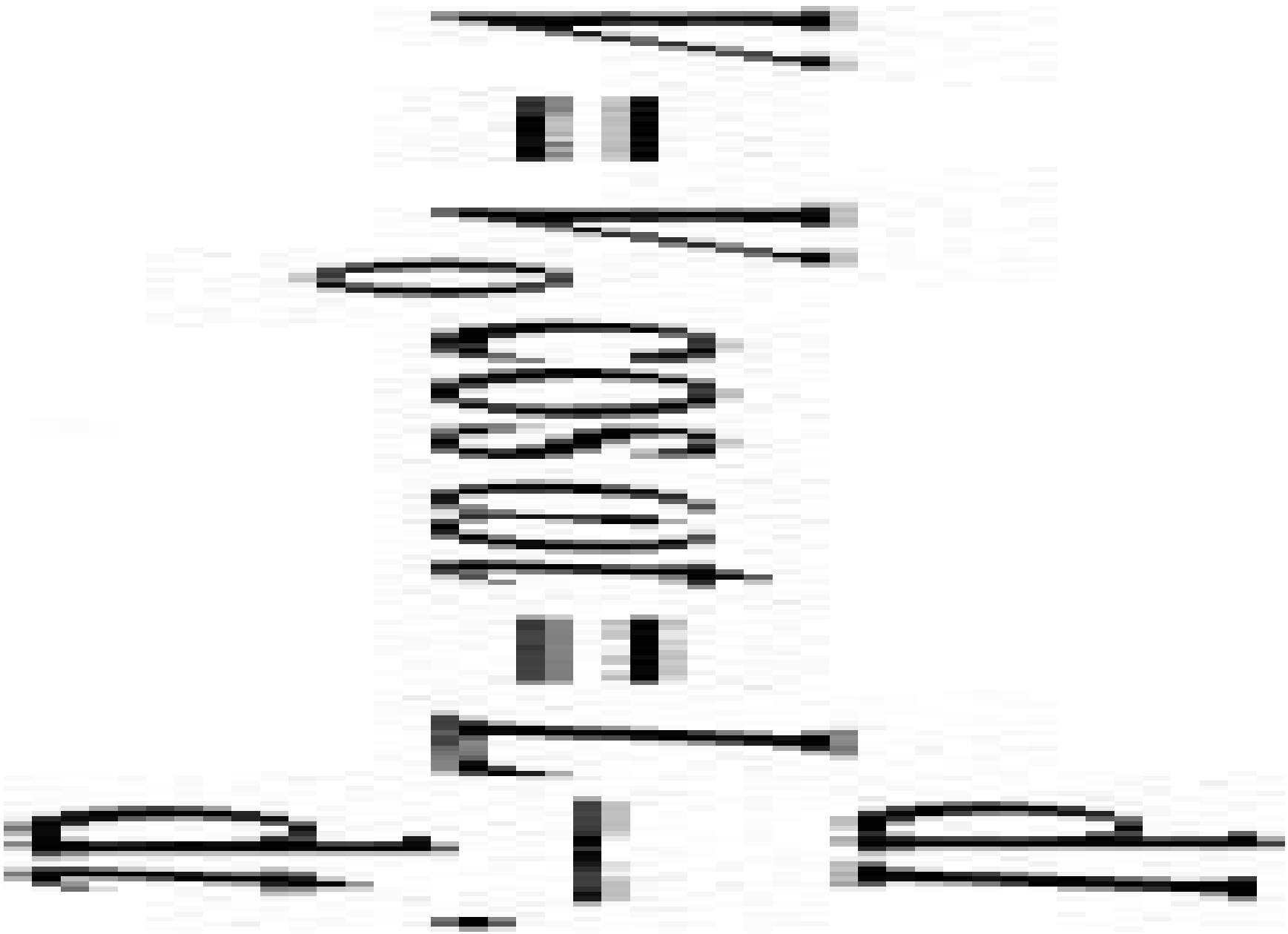


where I is the current flowing in the leads to the capacitor plates. Rearranging this equation,

(7.4)

The current is not in step with the voltage, and the phase of the current leads the phase of the voltage by $\pi/2$, as illustrated in Figure 7.3(b). Notice that the amplitude of the current given by Equation (7.4) has the same form as the amplitude in Equation (7.3) except that the resistance R is replaced by $(1/\omega C)$. The quantity $(1/\omega C)$, which has the same dimensions as R and is measured in ohms, is called the impedance of the capacitor. At low frequencies the impedance of a capacitor is large. For example, a capacitance of 1 μF (quite a large value) has an impedance to the 50 Hz mains supply of 3200 ohms: the amplitude of the current delivered by the generator is the same as if the capacitor were replaced by a 3200 ohm resistor. On the other hand at high frequencies, $1/\omega C$ becomes small, and capacitors then present a low impedance to the flow of alternating current.

Finally, in Figure 7.3(c), the circuit element across the generator is a coil with self-inductance L. This circuit is a rather unrealistic one, since coils usually have a resistance which is not negligibly small. However, it is convenient for the moment to imagine that such an idealized self-inductance exists, and then later to deal with circuits containing both inductance and resistance. The e.m.f. caused by the changing current in the inductance is L dI/dt. In the circuit of Figure 7.3(c) this e.m.f. balances the generator voltage, and



Integrating this equation, and setting the constant of integration equal to zero, since the circuit contains no source of steady current, we obtain

or

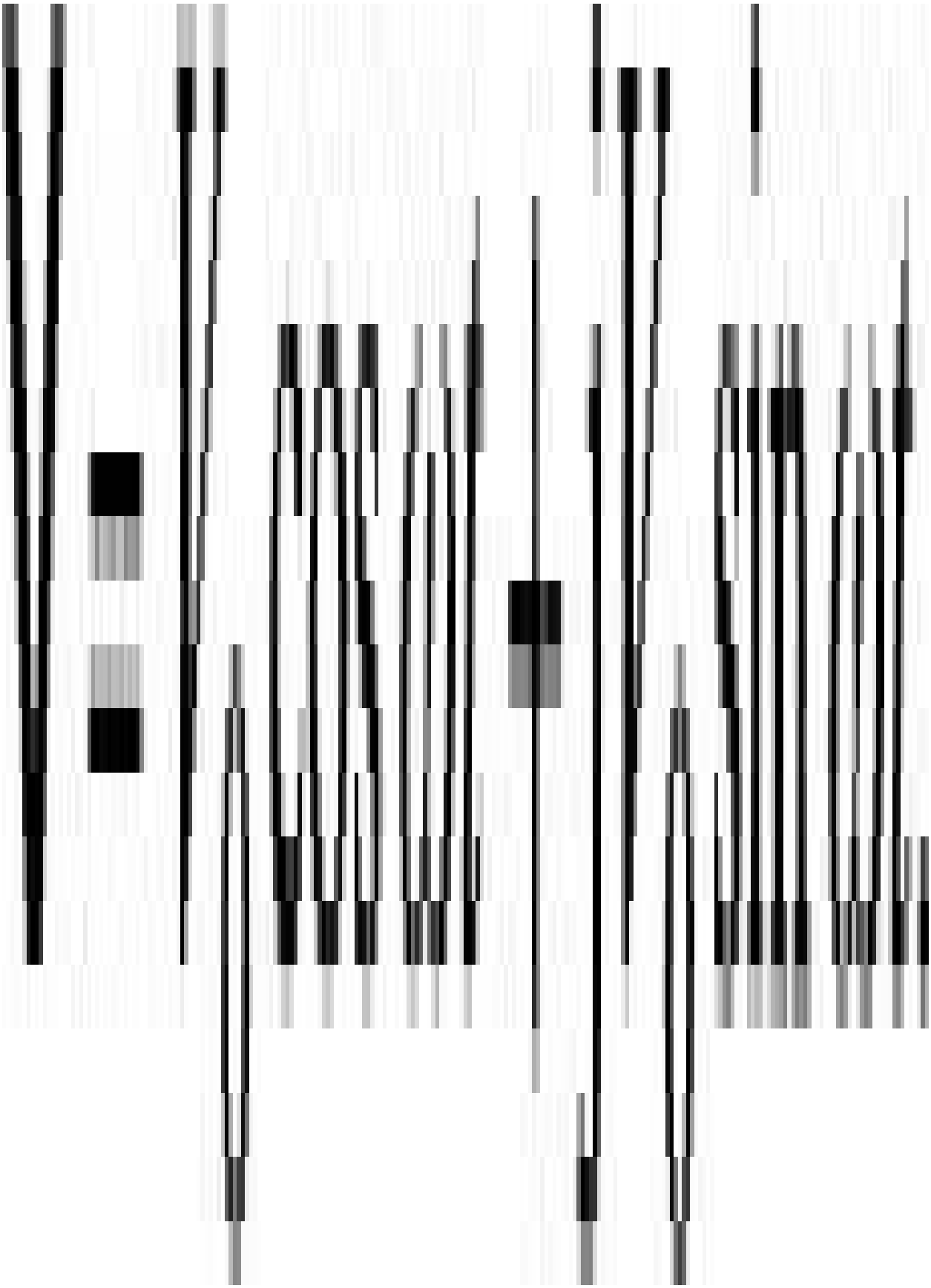
(7.5)

As with the capacitor there is a phase difference of π/2 between current and voltage, but in the coil the current lags behind the applied voltage. The amplitude of the current is V0/ωL, and here, by comparison with Equation (7.3), the resistance R is replaced by the impedance ωL of the inductance. For a coil with a self-inductance of 10 mH, the impedance to a 50 Hz mains supply is only about 3 ohms, but at high frequencies the impedance of a coil becomes large.

7.4 THE PHASOR DIAGRAM AND COMPLEX IMPEDANCE

At a particular moment, the value of an alternating voltage is specified by two quantities, namely its amplitude and phase. The phase is the argument of a sine or cosine function, that is to say it is a dimensionless quantity which can be envisaged as an angle. This is illustrated in Figure 7.4, which shows a triangle relating the instantaneous voltage $V_0 \cos \omega t$ to the amplitude V_0 and the phase angle ωt. Now imagine this triangle in the complex plane as shown in Figure 7.5. The side of length $V_0 \cos \omega t$ is placed on the real axis, and the bold arrow of length V_0 is a line joining the origin to the point with real coordinate $V_0 \cos \omega t$ and imaginary coordinate $V_0 \sin \omega t$. As time passes, the angle ωt increases uniformly, and this line rotates about the origin at constant angular speed ω. Its projection on to the real axis is $V_0 \cos \omega t$, equal to the instantaneous value of the alternating voltage. This is not surprising, since a simple A.C. generator consists of a coil rotating at constant angular speed, and the output voltage depends on the projection of the coil's area on to a plane perpendicular to the magnetic field. Figure 7.5 amounts to a plan view of such a rotating coil.

Because it illustrates the phase of the alternating voltage, the diagram in Figure 7.5 is called a phasor diagram, and the bold arrow on the diagram is itself called a phasor. The complex number specifying the coordinates of the head of the arrow is



In this equation j stands for the imaginary number the symbol j rather than i is commonly used in electromagnetism. We have written the complex number V in bold type. This type is different from the bold type used for three-dimensional vectors (it is a sanserif type with no tails on the letters—compare V and V). In practice complex numbers and three-dimensional vectors do not occur together in the same equation, and it is always clear what the bold symbols mean.

Figure 7.4. The phase angle ωt relates the amplitude V0 of the voltage to its instantaneous value V0 cos ωt.

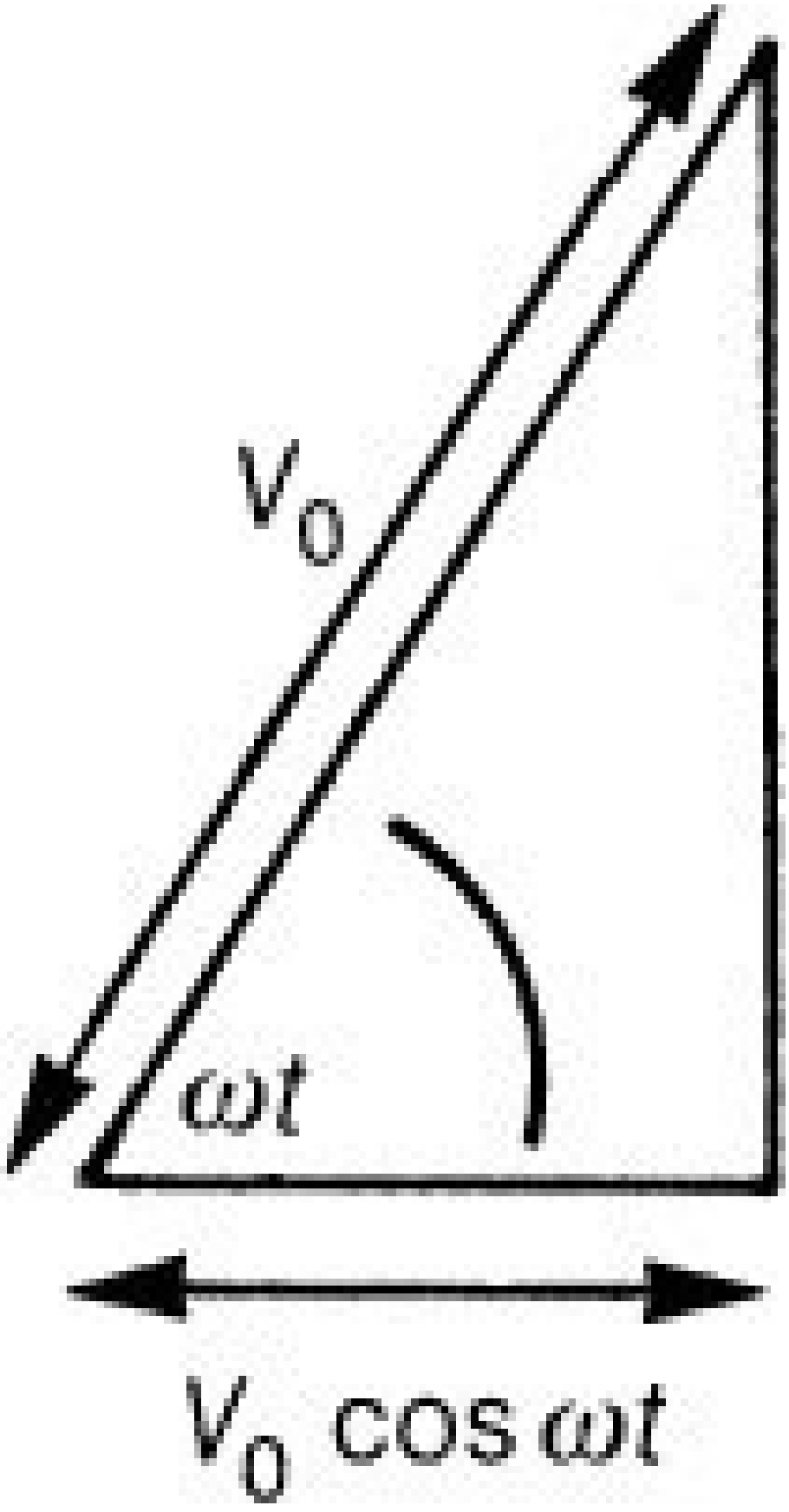
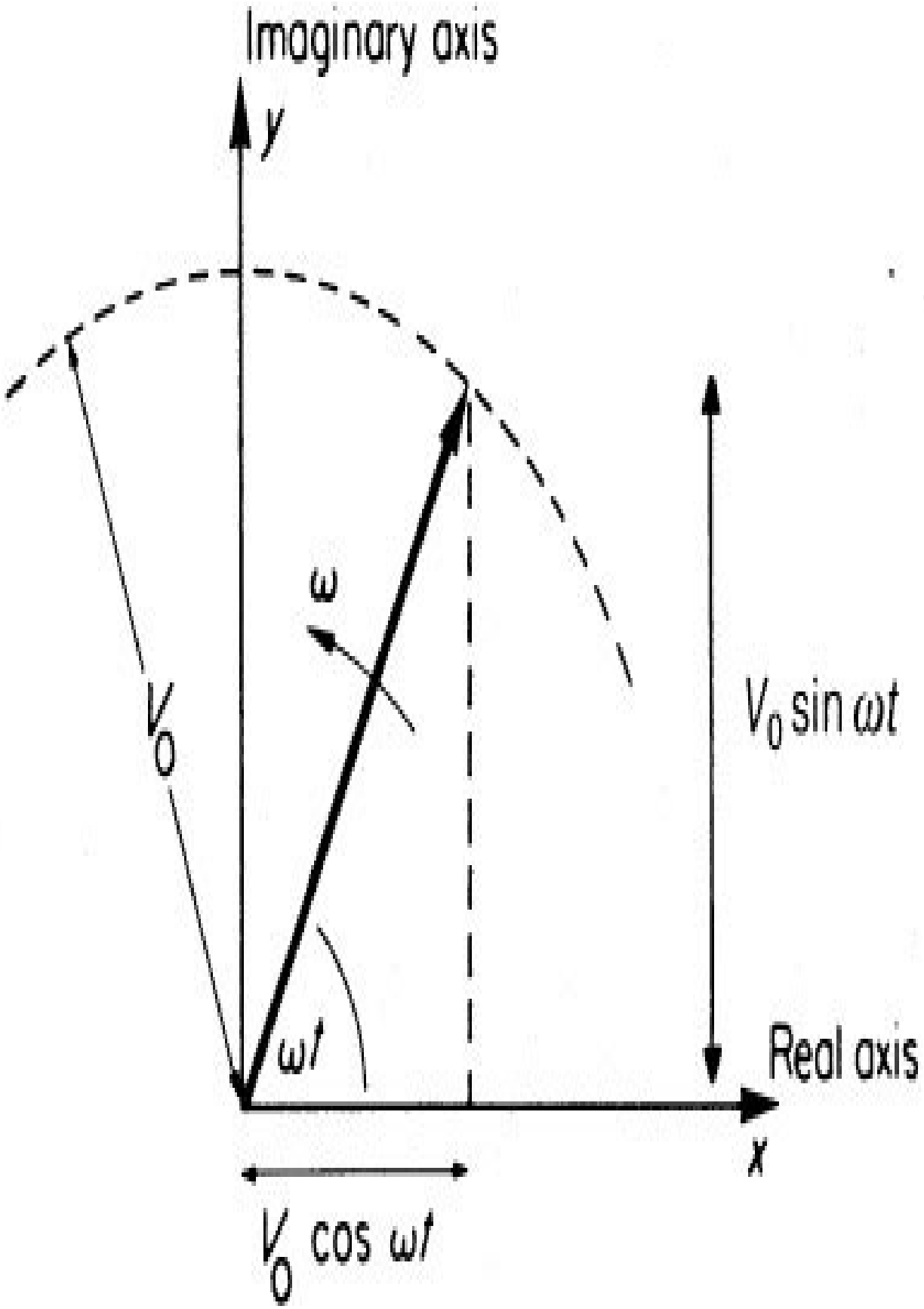
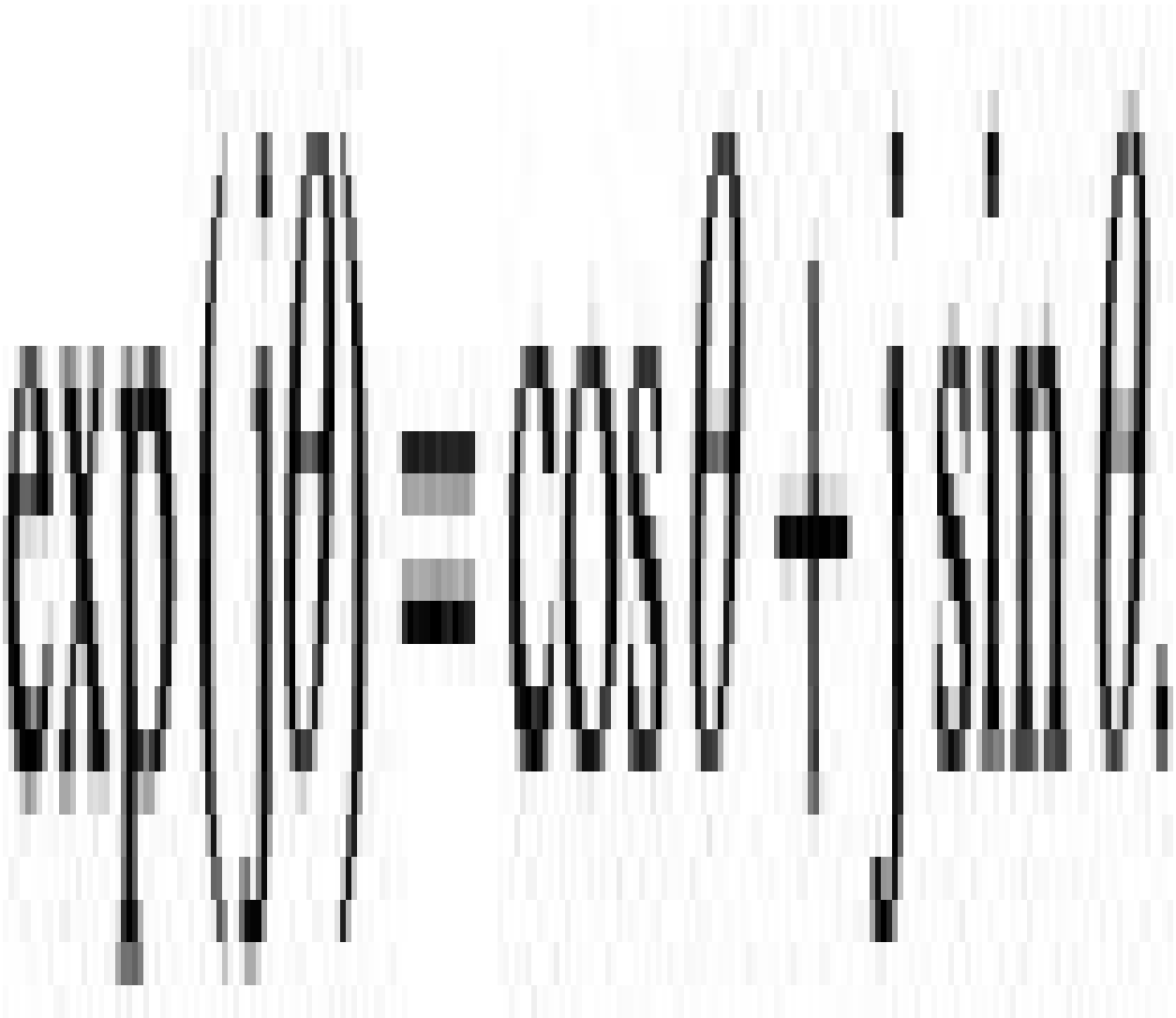

V_0
ωt
$V_0 \cos \omega t$

Figure 7.5. The representation of an alternating voltage in the complex plane.

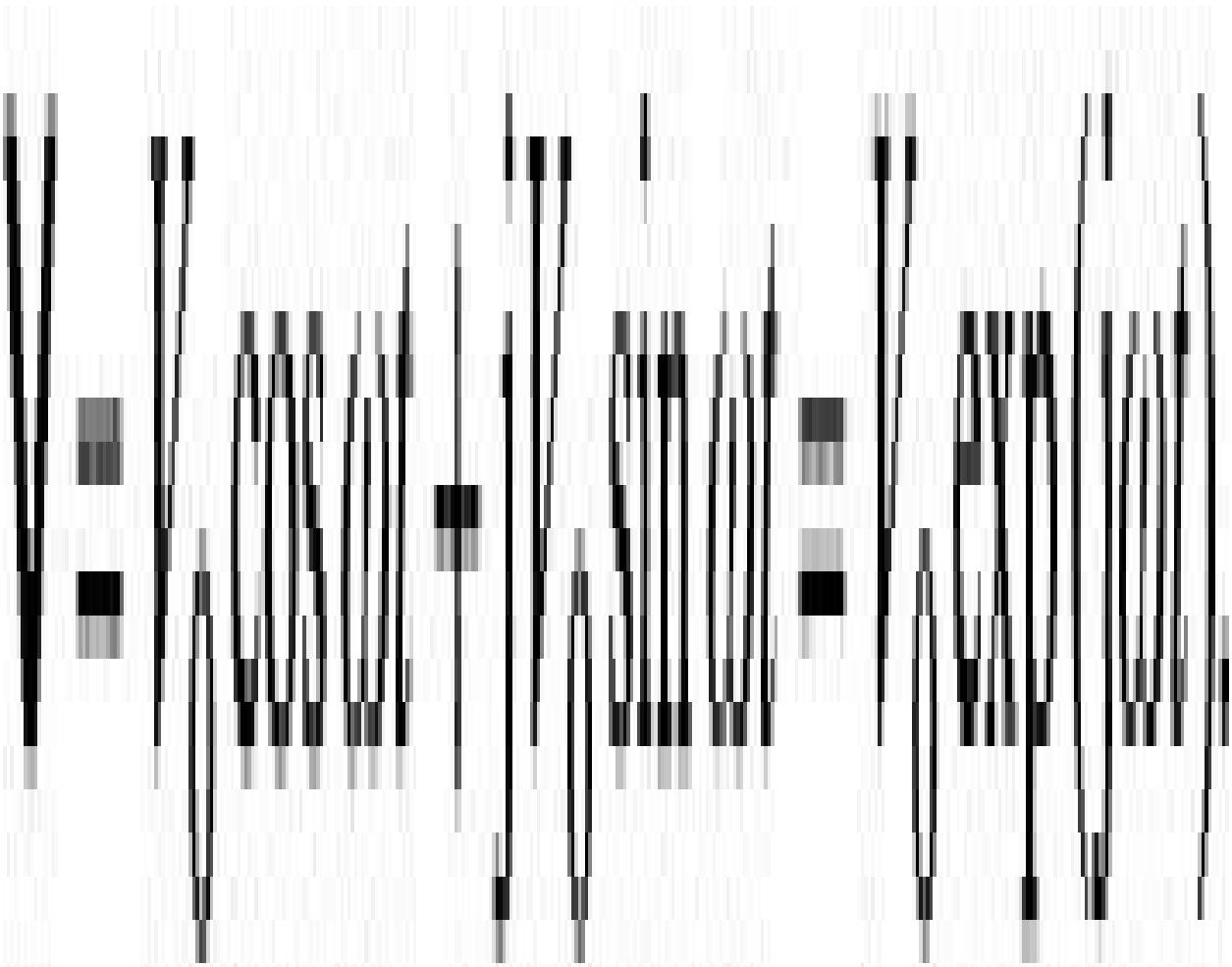

Imaginary axis
y
ω
V_0
$V_0 \sin \omega t$
ωt
Real axis
x
$V_0 \cos \omega t$

The phase may be expressed in a concise way with complex numbers by using Euler’s formula, which states that



Applying this formula to the complex number V, we find that

(7.6)



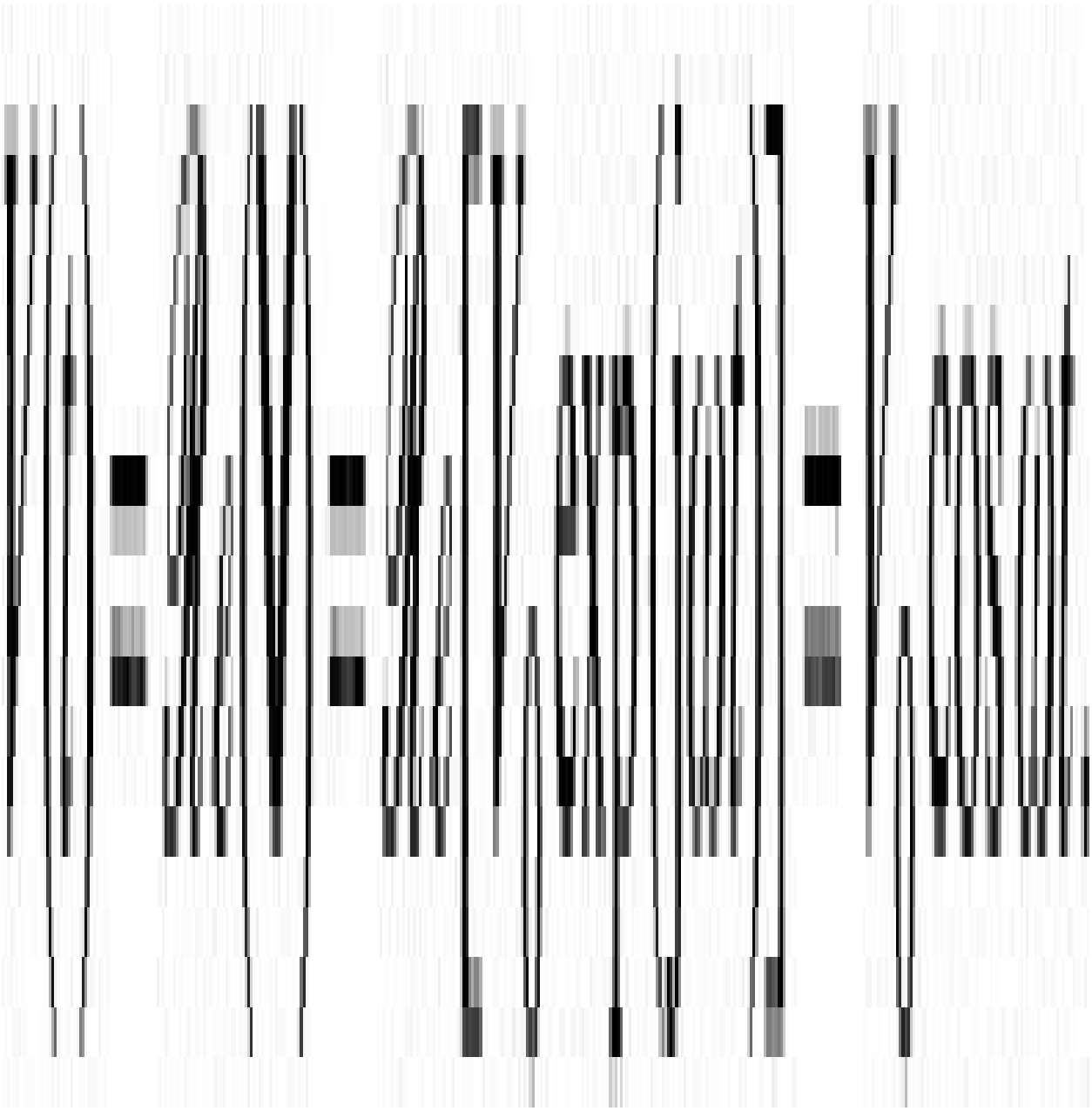
The modulus of the complex number exp (jωt) is one, and it represents the coordinates of the head of a unit phasor, i.e. a phasor of unit length rotating in the complex plane.

When using phasors and complex numbers for alternating voltages one must always remember two things:

(a) phase is defined with respect to some reference value, in our example by choosing t = 0 when the voltage has its maximum value V0, and

(b) only the real part of the complex number represents a physical quantity, and the imaginary part has no physical meaning. The actual voltage V(t) at time t is

(7.7)



At first it may seem as if the introduction of complex numbers has led to a lot of complication for little benefit. However, complex numbers do in fact greatly simplify calculations concerning circuits in which phase differences occur and which may contain many components. We shall start by using phasors and complex numbers to describe the phase differences between voltage and current in single components, and then go on to consider some simple circuits containing more than one component.

Alternating currents as well as voltages may be represented by phasors, though of course because they have different dimensions from voltages, they require a separate scale for diagrams in the complex plane. The current phasors rotate at the same angular speed as the voltage phasors, but may lead or lag by a constant phase angle. This is illustrated for an ideal inductance in Figure 7.6(c), in which the phasors are frozen at the moment when the current phasor points along the real axis. According to Equation (7.5), the amplitudes of voltage and current for a self-inductance L are related by $V0 = I0 \times \omega L$, and the current lags behind the voltage by the phase angle $\pi/2$. The voltage phasor is thus perpendicular to the current phasor. At the moment when the current phasor is real, as in Figure 7.6(c), the voltage phasor is purely imaginary. Figures 7.6(a) and 7.6(b) are the corresponding diagrams for resistance and capacitance.

The phase relations illustrated in the phasor diagrams in Figure 7.6 can be expressed in a very simple way using complex variables. Rewriting Equation (7.3) in a complex notation, we have for the current through a resistance R

$$I = \frac{V_0}{R} \exp(j\omega t)$$

$$= V/R$$

or

(7.8)

In this equation relating complex variables, the resistive impedance R is real.

Equation (7.4) for the current in the capacitative circuit becomes, when written in the complex notation,

$$I = \omega C V_0 \exp\left[j\left(\omega t + \frac{\pi}{2} \right) \right]$$

$$= \omega C V \exp\left(\frac{j\pi}{2} \right)$$

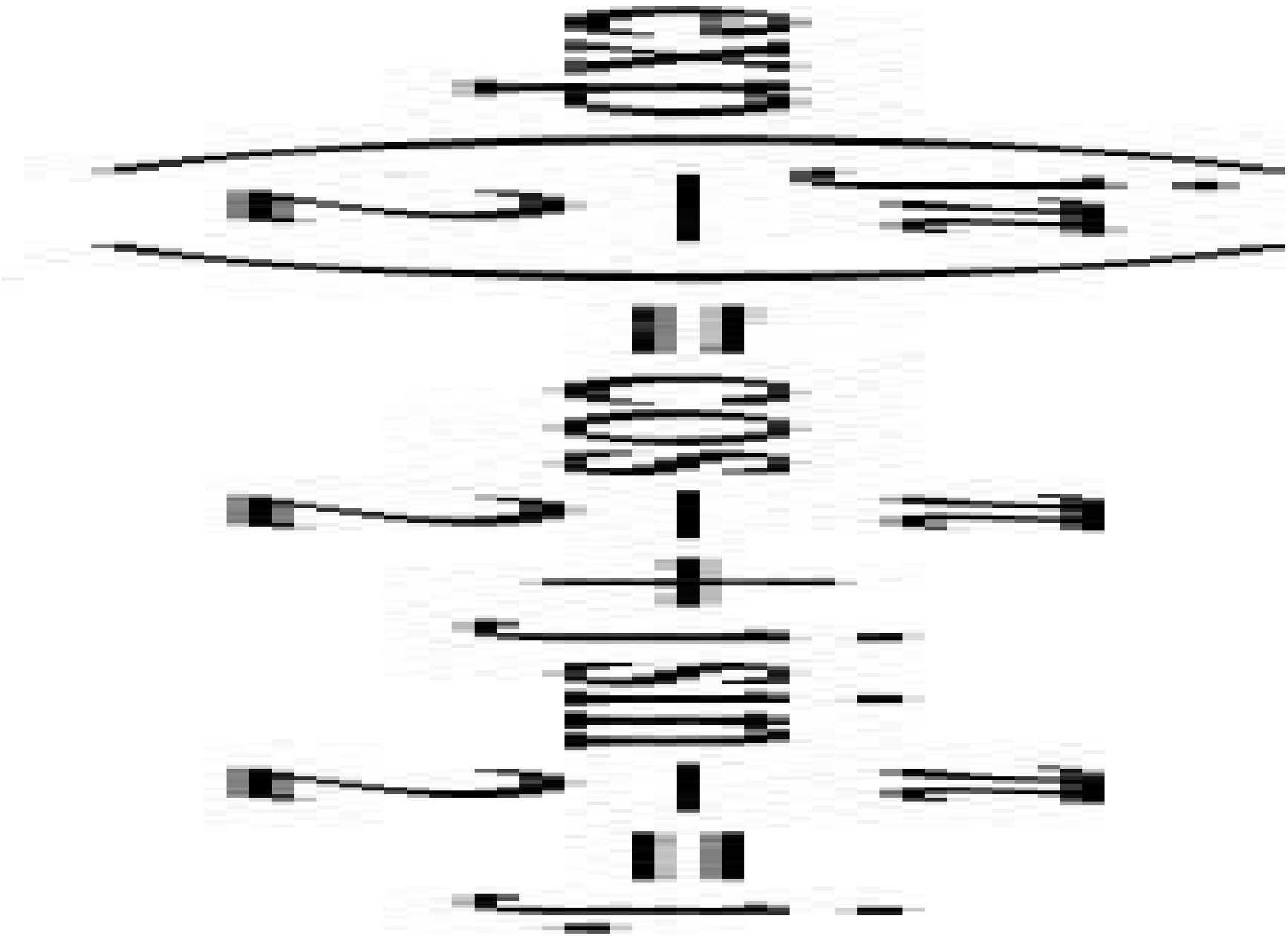
or

(7.9)

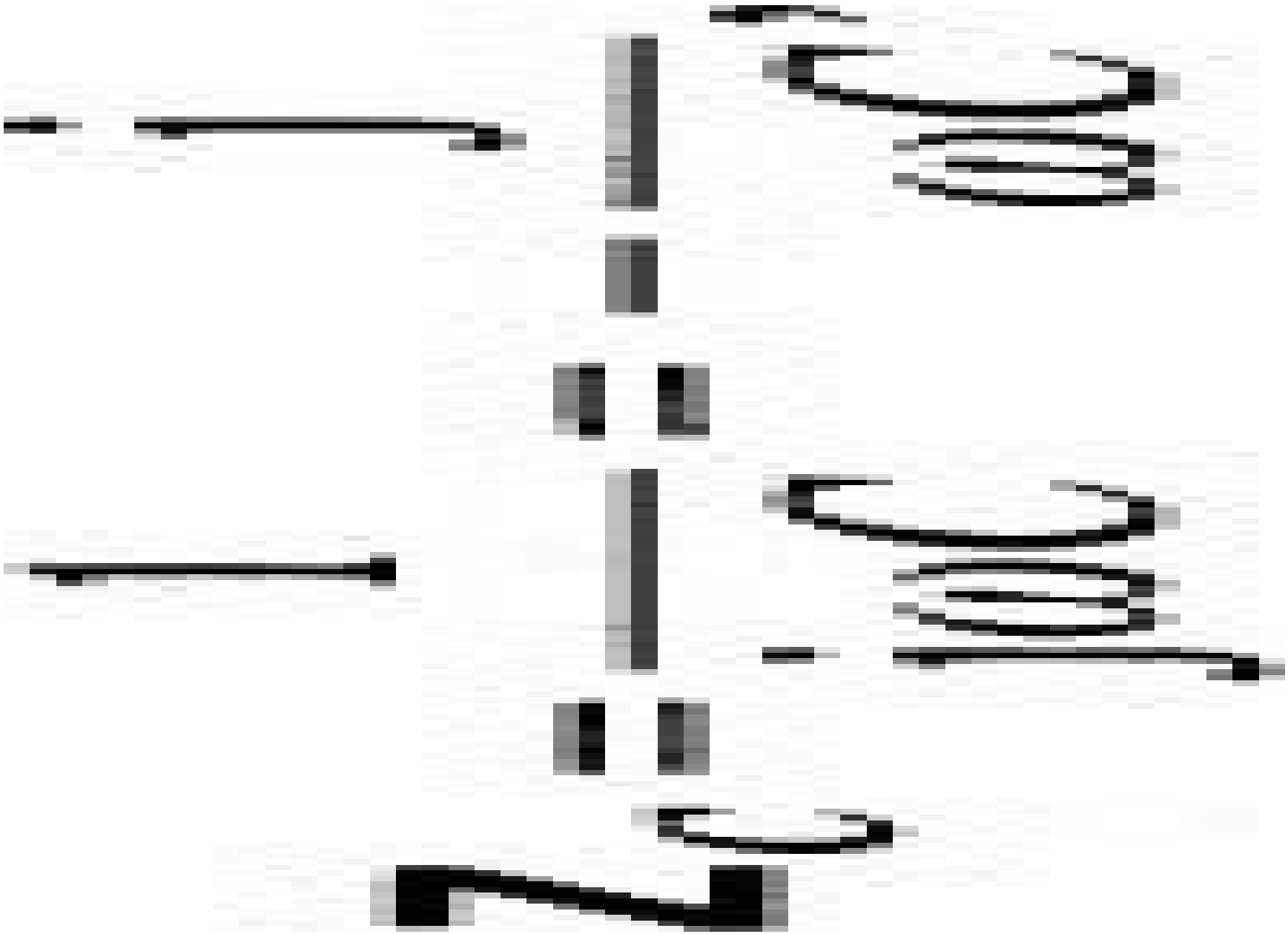
Figure 7.6. Phasor diagrams.

(a) Resistance (b) Capacitance (c) Inductance

since

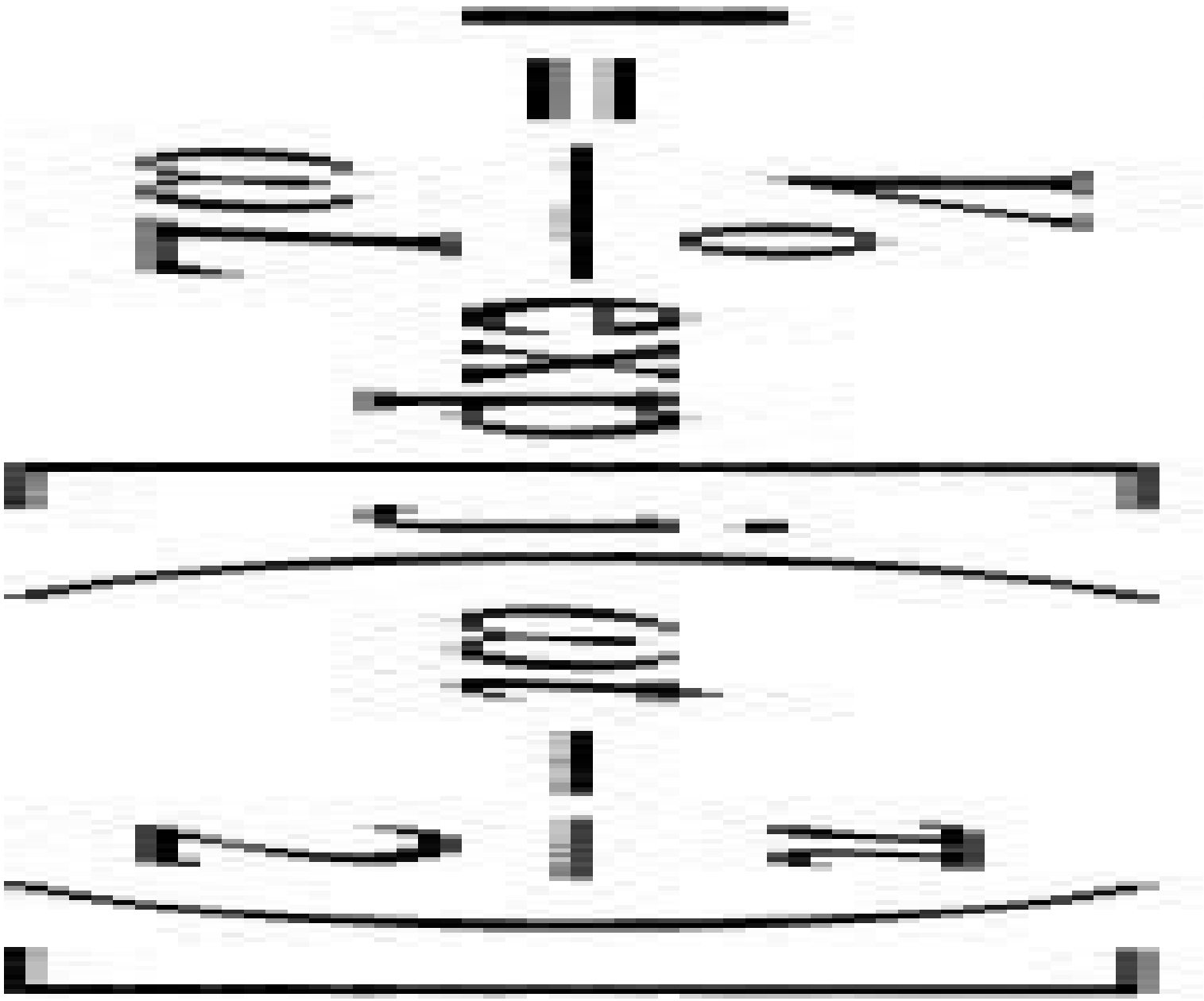


Multiplication by (–j) is equivalent to the rotation of a phasor through ($-\pi/2$), and Equation (7.9) tells us both the amplitude and phase of the voltage in terms of the current. This equation has the same form as Ohm's law if we assign to the capacitor a complex impedance



so that Equation (7.9) becomes V = IZC. Because the current phasor has been chosen to lie along the real axis, the point representing the impedance ZC in the complex plane lies on the voltage phasor along the negative imaginary axis. As they are labelled in Figure 7.6, the light arrows are measured in volts, and indicate phasors. However if you prefer you can rub out the I0 on the label, re-scale in ohms instead of volts, and regard the light arrows as representing complex impedances. The impedance has no time dependence, of course, and the impedance diagrams do not rotate.

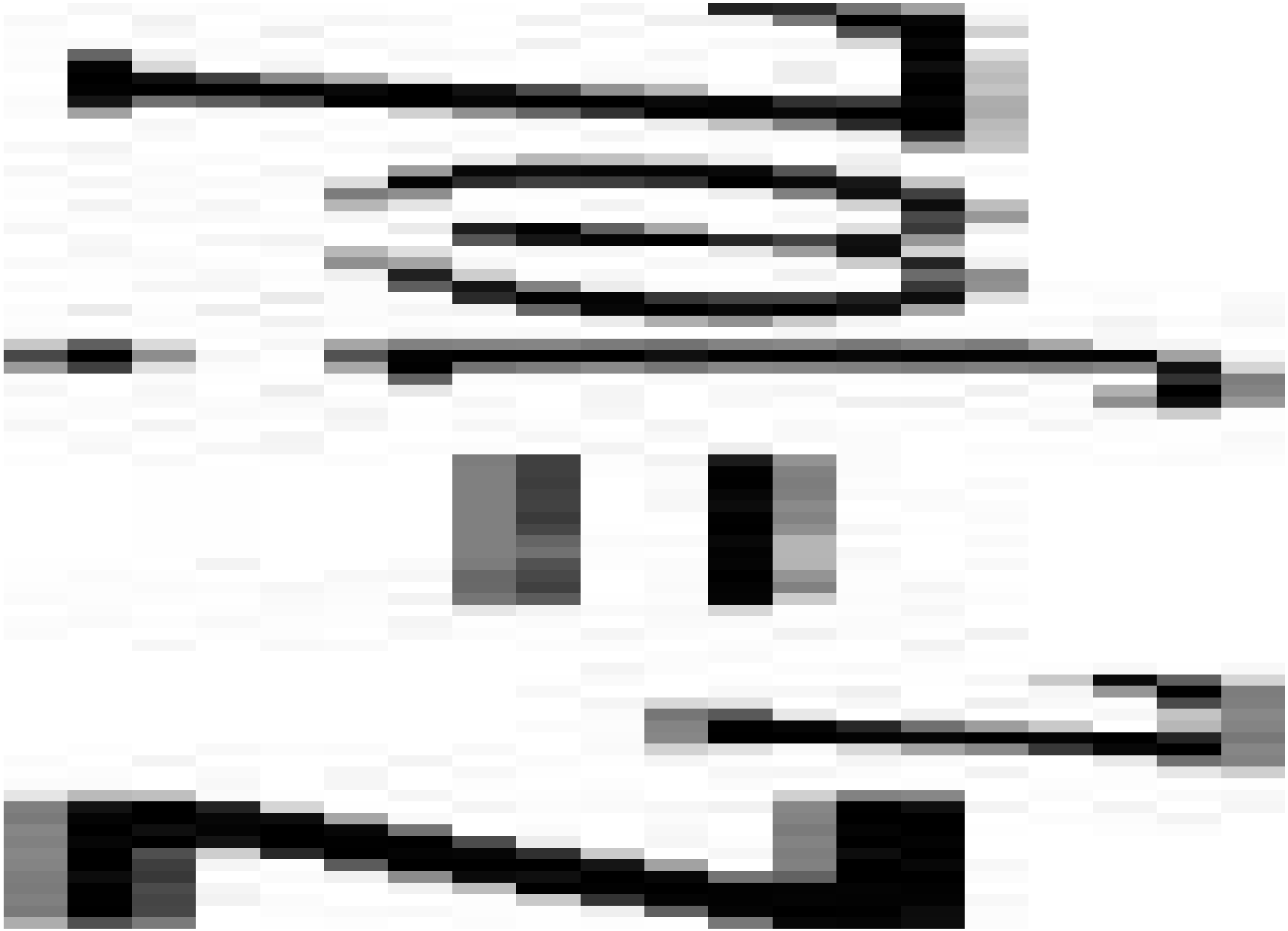
Finally, Equation (7.5) for the current through an inductance is written as



or

(7.10)

The inductance has a complex impedance



represented by a vector pointing upwards, along the positive imaginary axis.

The purely imaginary impedances we have derived apply only to idealized components. Real capacitors have a leakage resistance between their plates, and the windings of real coils have a non-zero resistance: in both instances the effect of the resistance is to reduce the phase angle between voltage and current to less than $\pi/2$. For most practical purposes capacitors closely approach the ideal, and in the analysis of an A.C. circuit it is usually adequate to represent a capacitor by an idealized component with impedance $1/j\omega C$. With coils however, it is often not possible to ignore the resistance of the windings.

Let us evaluate the complex impedance of a coil with a resistance R to direct current, when an alternating voltage of angular frequency ω is applied across it. At very high frequencies the behaviour of the coil is complicated because in addition to the inductive and resistive effects, it becomes necessary to take into account the capacitance between neighbouring turns on the coil. However, we shall restrict ourselves to frequencies at which the impedance of this small capacitance is extremely large compared with the inductive impedance of the coil. There is then no significant current carried by the capacitance, and it is a good approximation to assume that at any moment the current I is the same throughout the windings of the coil. Furthermore, the ‘skin effect’, which causes high frequency currents to concentrate near the surface of a conductor*, is not important, and the resistance of the coil is the same as for direct current.

At moderate frequencies, then, the current is uniform in the windings, and the coil has a self-inductance L = (flux/current) which is independent of frequency. This inductance generates an e.m.f. $L\,dI/dt$ opposing any change in current. The external e.m.f. V driving the current must balance this e.m.f. and also account for the additional voltage drop IR due to the resistance. Hence

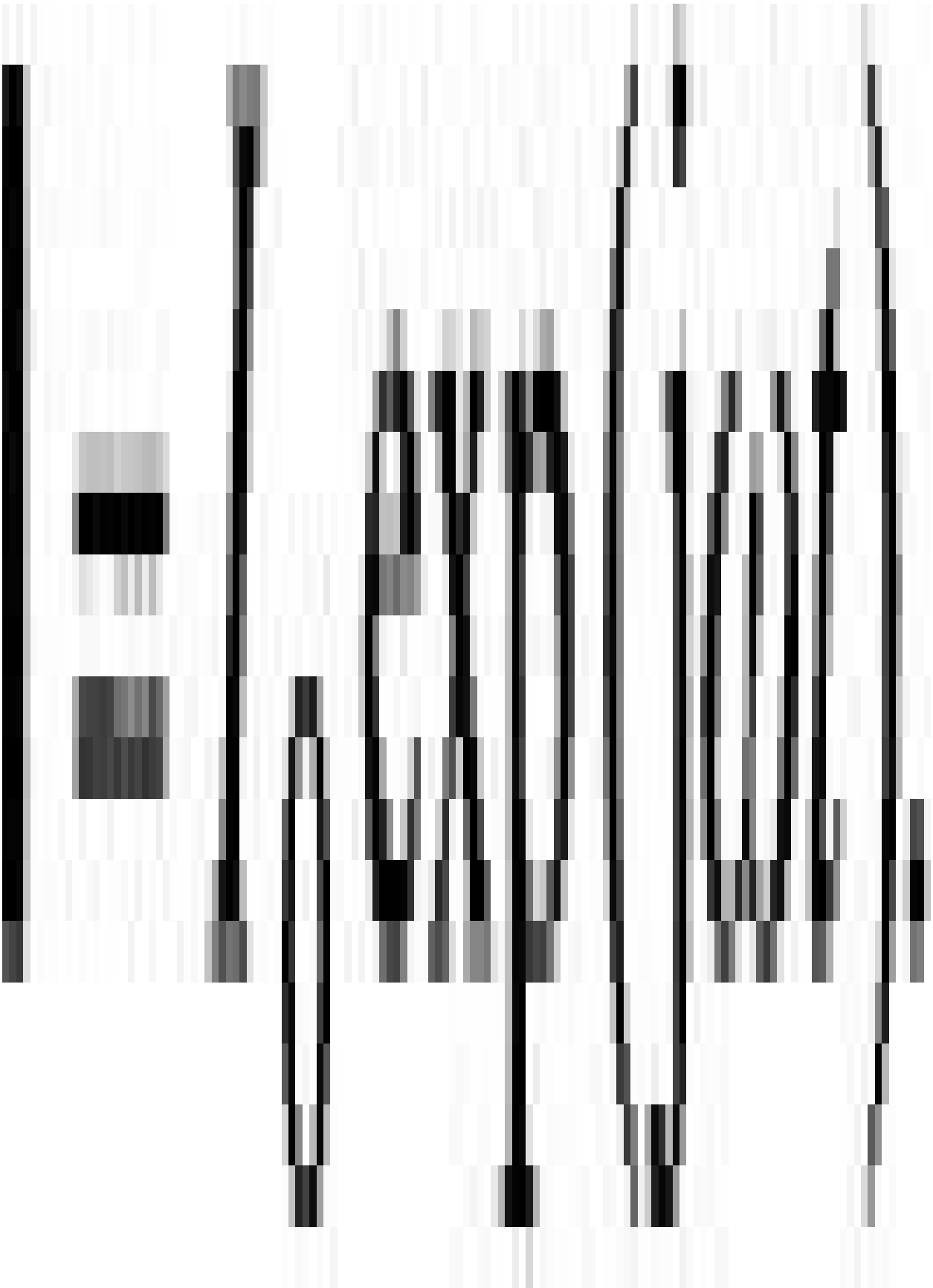
(7.11)

We can imagine that the coil is replaced by a circuit consisting of an ideal inductance L in series with a separate resistance R, as shown in Figure 7.7. The voltage and current in this circuit satisfy Equation (7.11) at all frequencies. At the moderate frequencies in which we are interested, it is an equivalent circuit which we may use to represent the real coil.

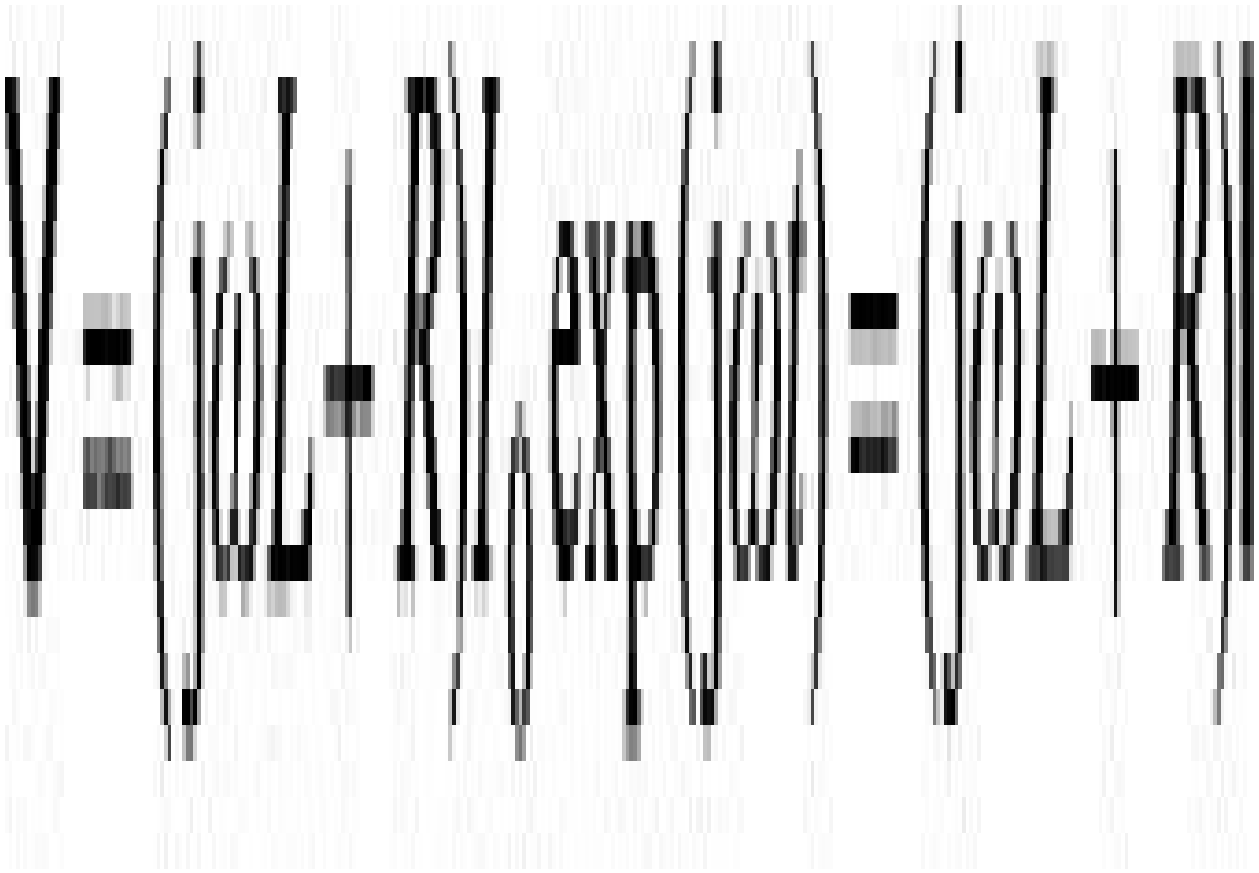
Now we must express Equation (7.11) in terms of complex numbers. To do this, all that is needed is to replace the actual voltage V and the current I by the complex numbers V and I:

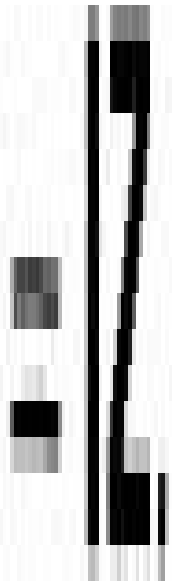
(7.12)

The real part of this equation is just Equation (7.11) once more. The imaginary part has no physical meaning, but is nevertheless useful because it allows us to introduce a complex impedance which again makes the equation look like Ohm's law. If we choose the time-zero when the current has its maximum value, then in the complex notation



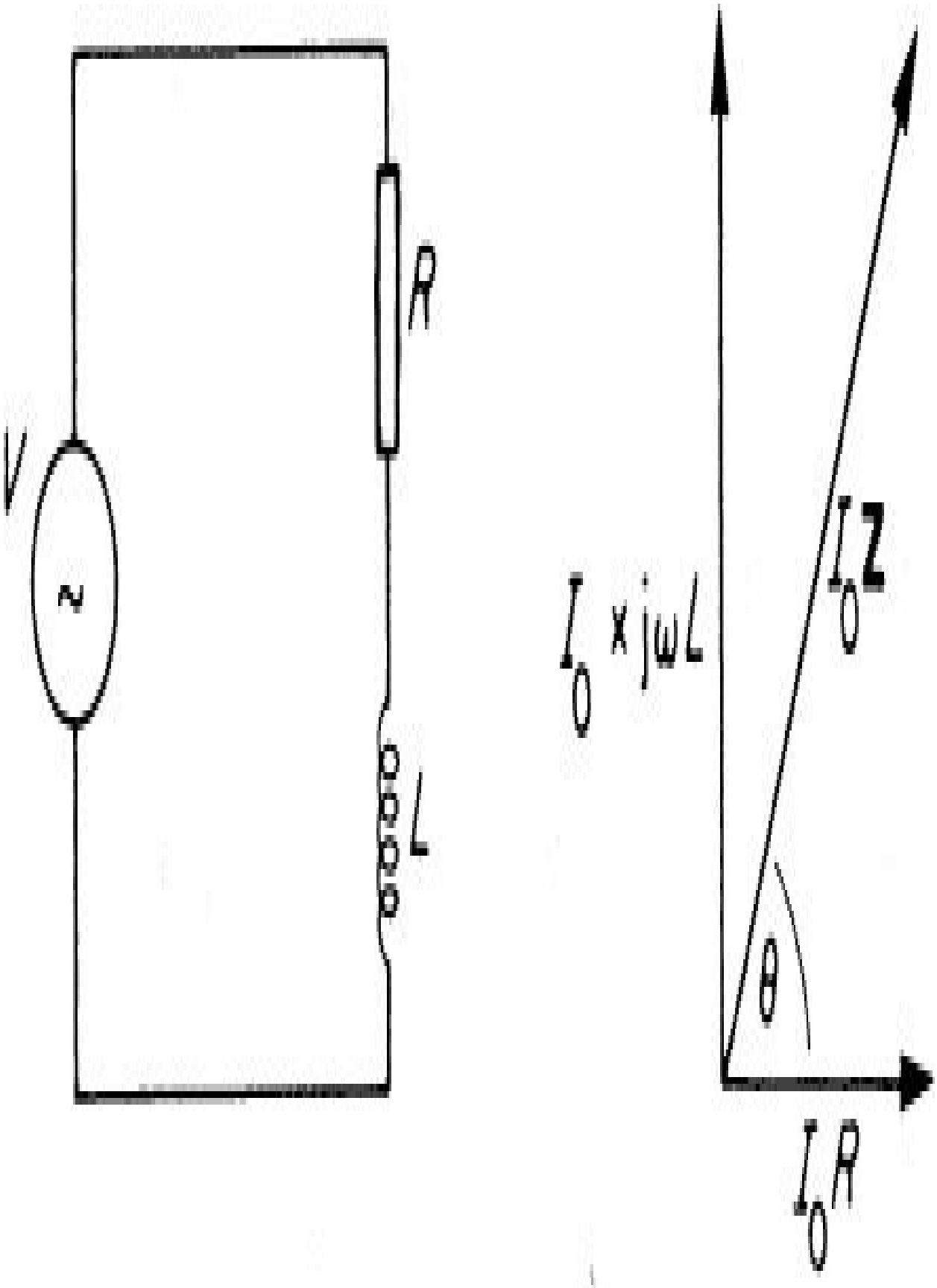
Substituting in Equation (7.12)





The complex impedance of the coil is $Z = (j\omega L + R)$, simply the sum of the impedances of the two components of the equivalent circuit. In the phasor diagram in Figure 7.7, the voltage V is represented by the vector sum of the voltages across L and R. From this diagram we see that the amplitude of the voltage is

Figure 7.7. Equivalent circuit and phasor diagram for an A.C. generator driving current through a coil.


R
V
~
L
$I_0 \times j\omega L$
$I_0 Z$
θ
$I_0 R$

$$V_0 = I_0\sqrt{R^2 + \omega^2 L^2},$$

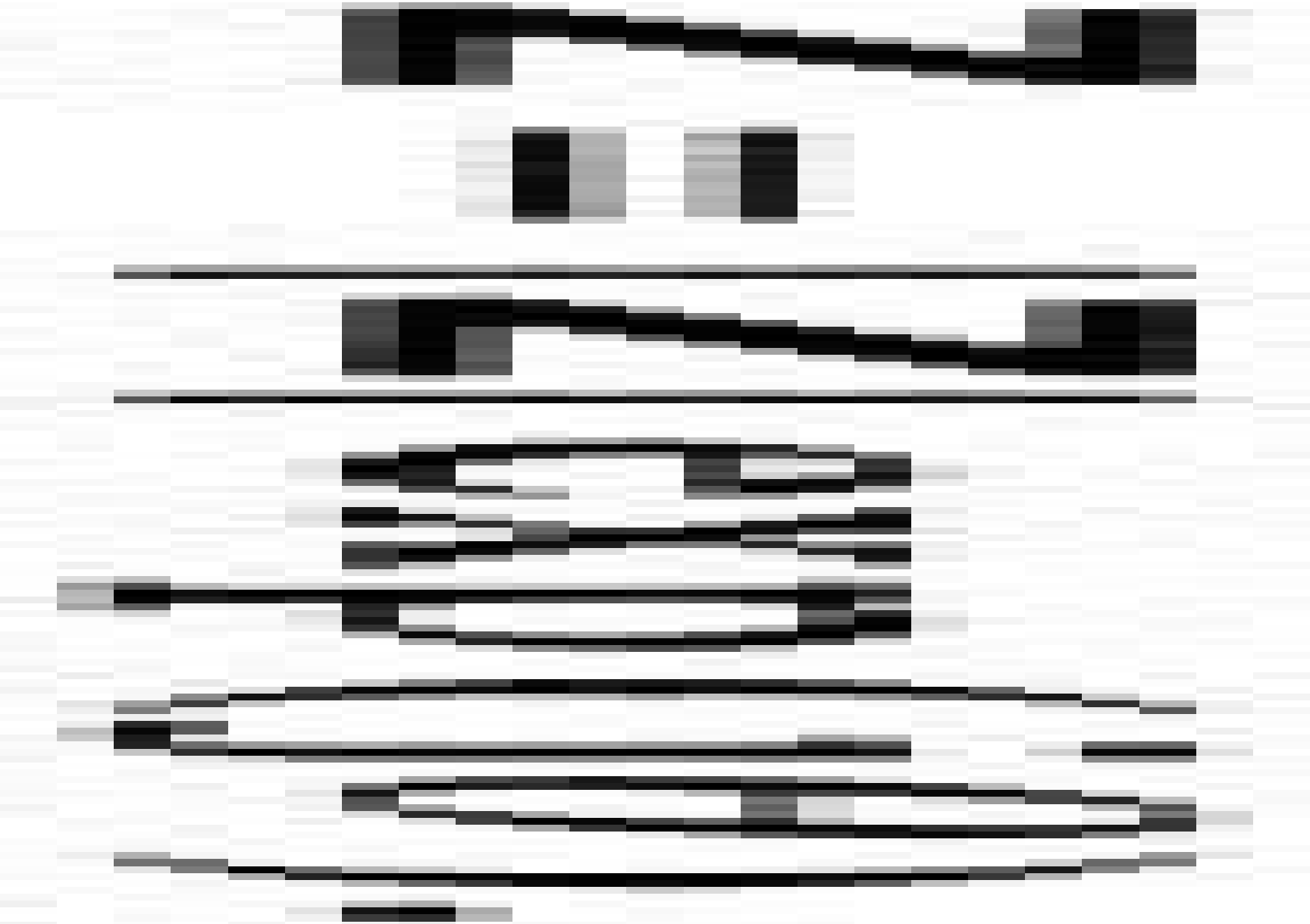
or, since $|Z|^2 = ZZ^* = (j\omega L + R)(-j\omega L + R) = \omega^2 L^2 + R^2$,

$$V_0 = I_0|Z|.$$

The voltage leads the current by the phase angle

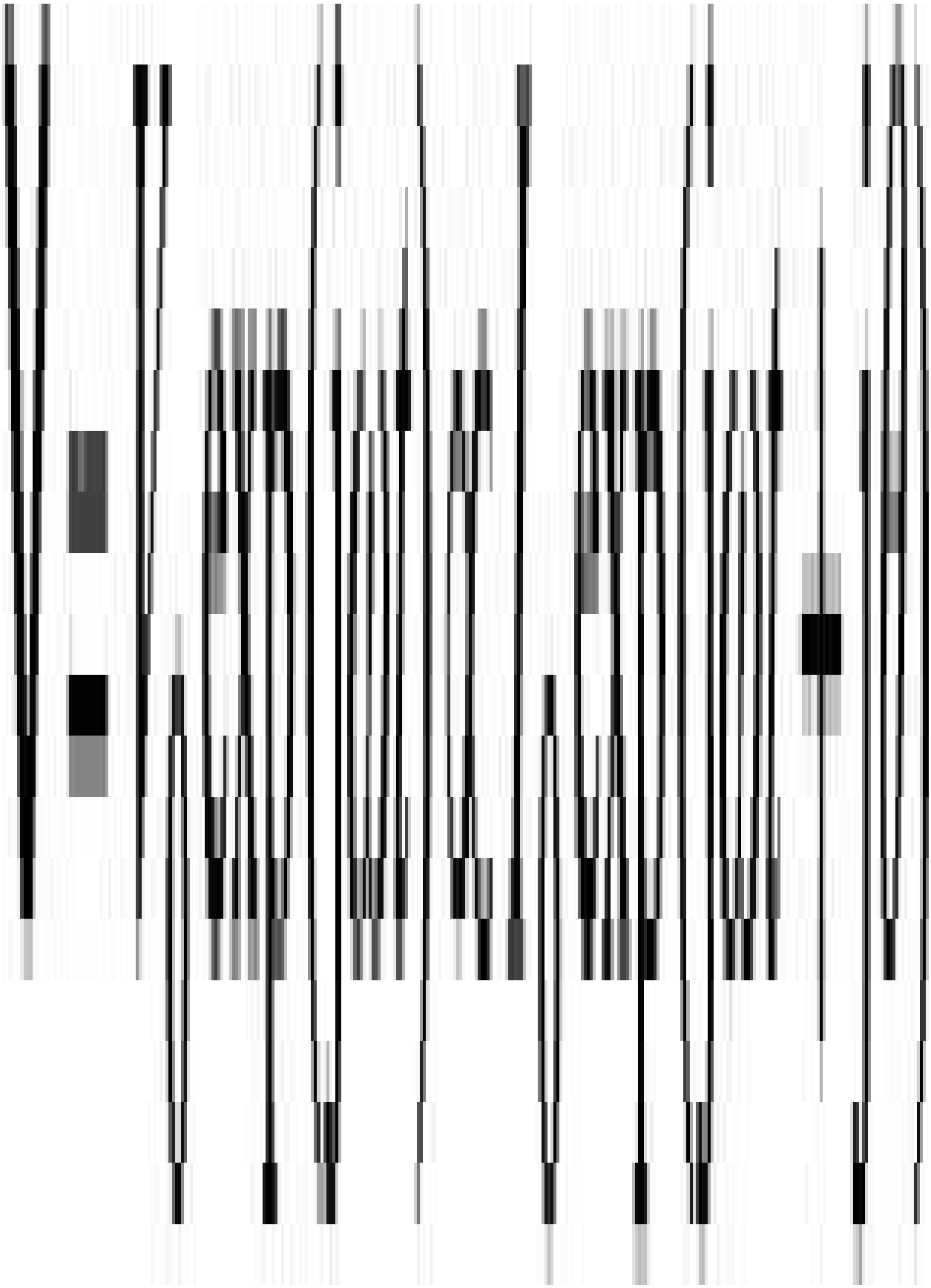
(7.13)

and the complex impedance can also be written in exponential form as

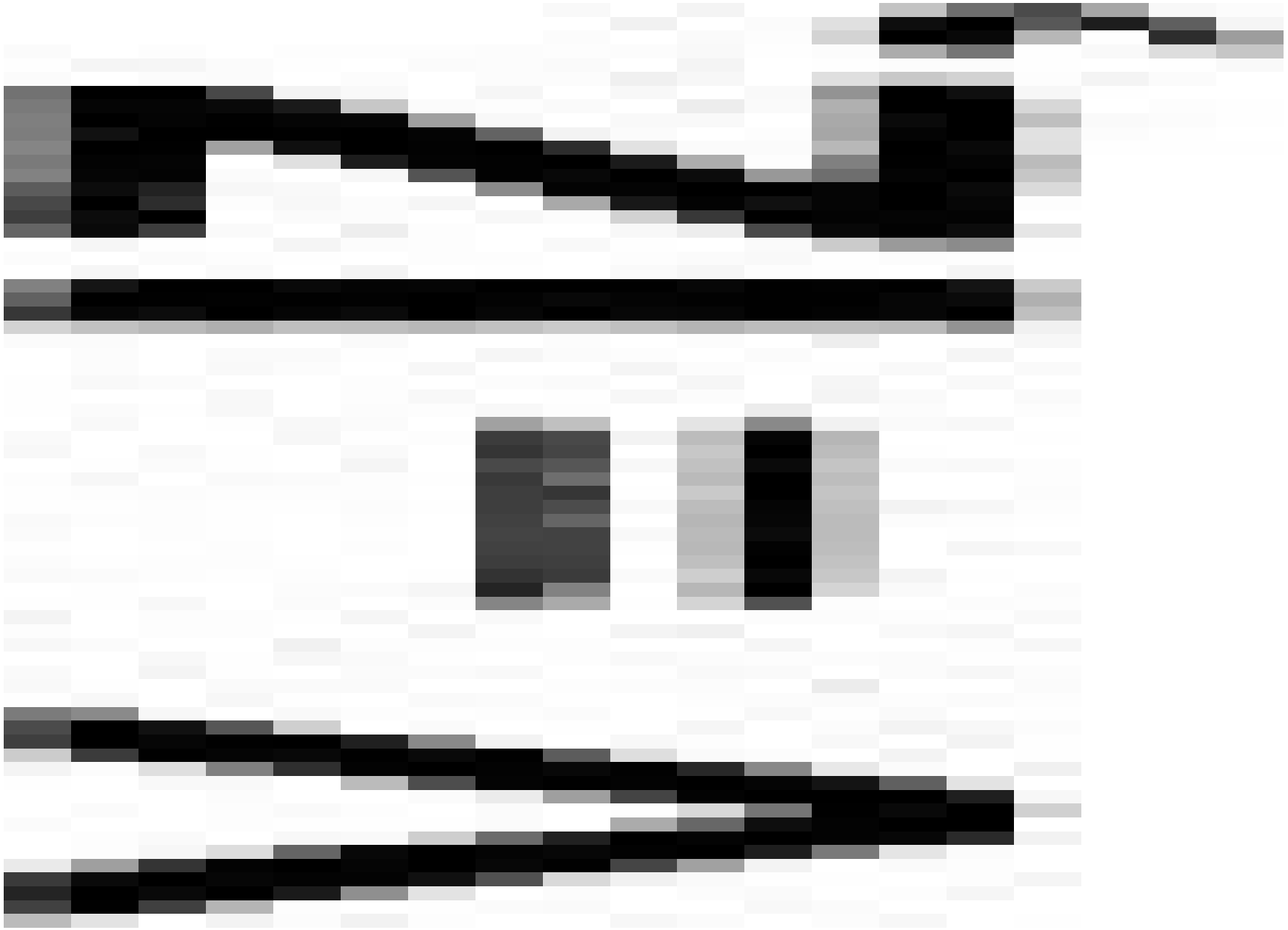


For typical coils used as smoothing chokes at audio frequencies in domestic equipment, the ratio of self-inductance to resistance is of the order of 0.05 henries/ohm. At the main supply frequency, ω 300 rads s−1 and the phase angle θ tan−1(15) 86°. Even at this low frequency the voltage and current are almost 90° out of phase. Nevertheless, as we shall see in section 7.6, the resistance of a coil sometimes plays an important part in the behaviour of A.C. circuits.

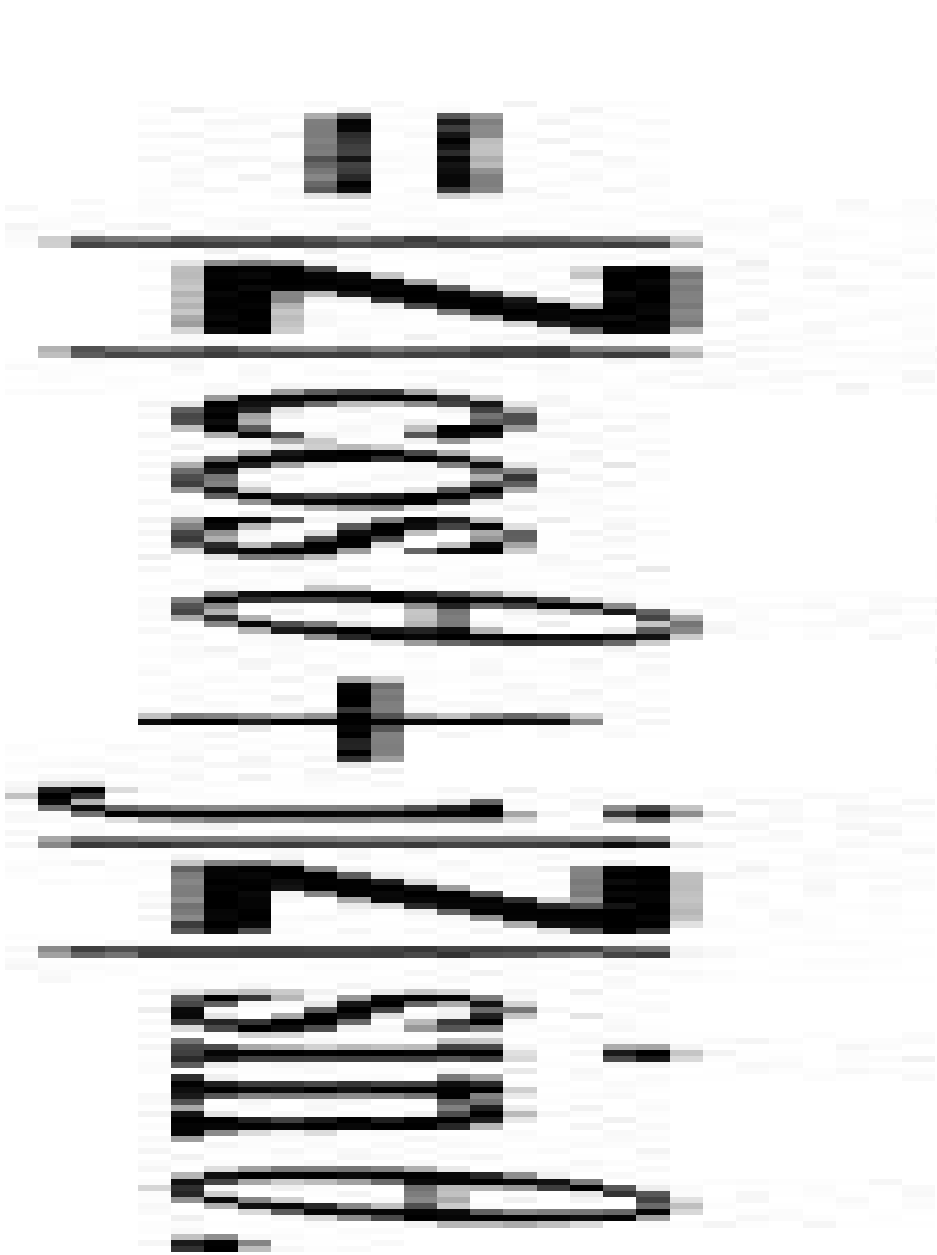
We have now obtained expressions for the impedances of resistors, capacitors and coils, which are all linear circuit elements in the sense that the amplitude of the current they carry is proportional to the amplitude of the voltage applied across them. They are called passive elements to distinguish them from active elements like dynamos or batteries which can generate e.m.f.s. When an alternating e.m.f. drives a current through a load made up of a network of passive elements, each of which is linear, then the response of the whole network must also be linear. We can define a complex impedance for such a load* in the same way as for the single circuit elements. The amplitude I0 of the current delivered by the generator is proportional to its voltage amplitude V0, but the current and voltage are in general not in phase. If the voltage leads the current by the phase angle θ, we can write



or



where



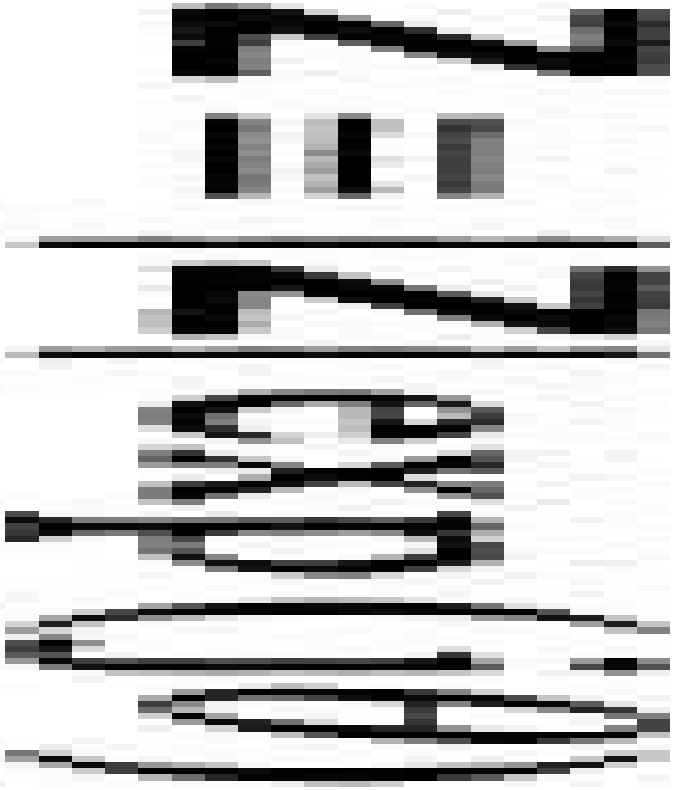
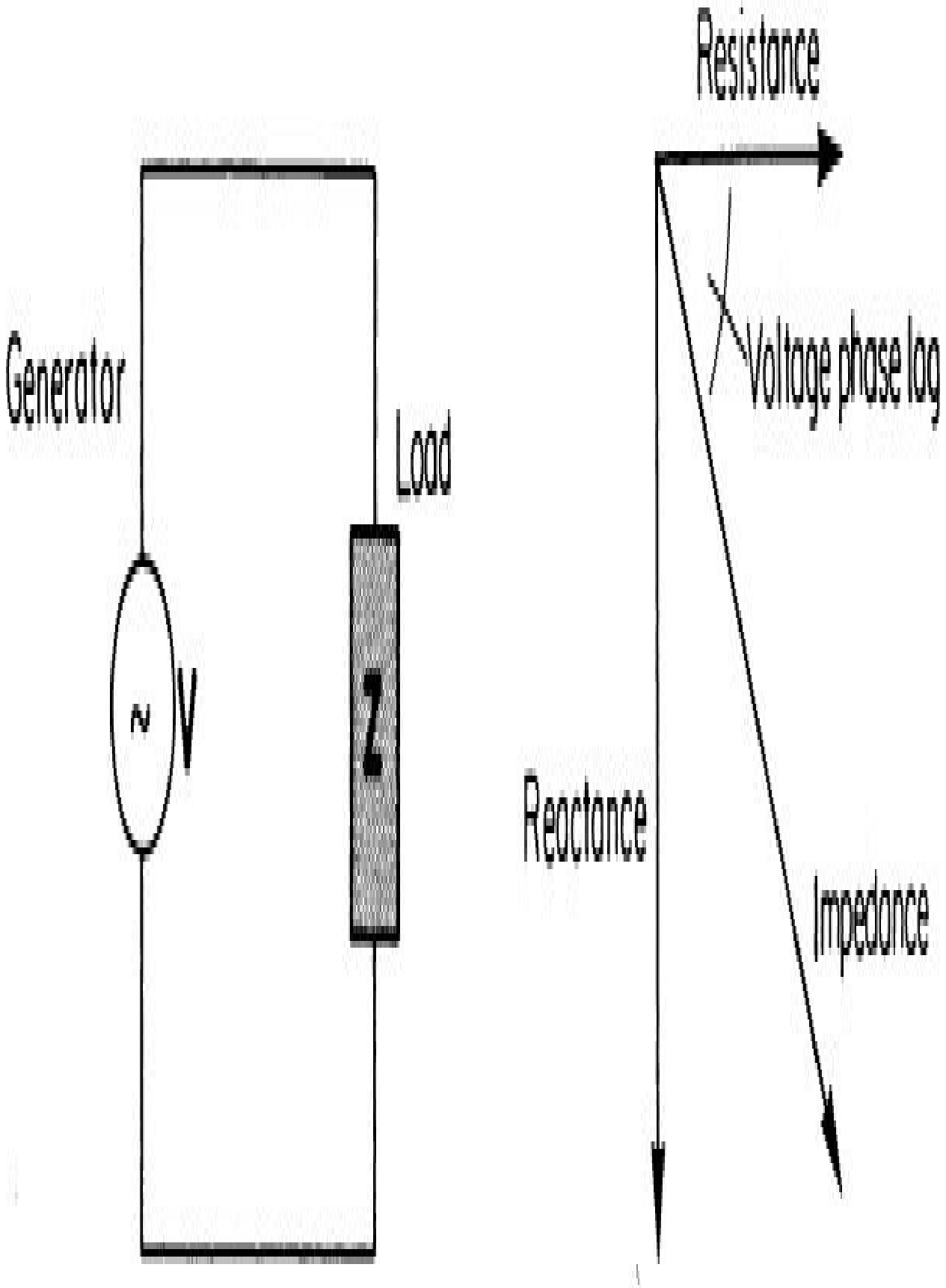


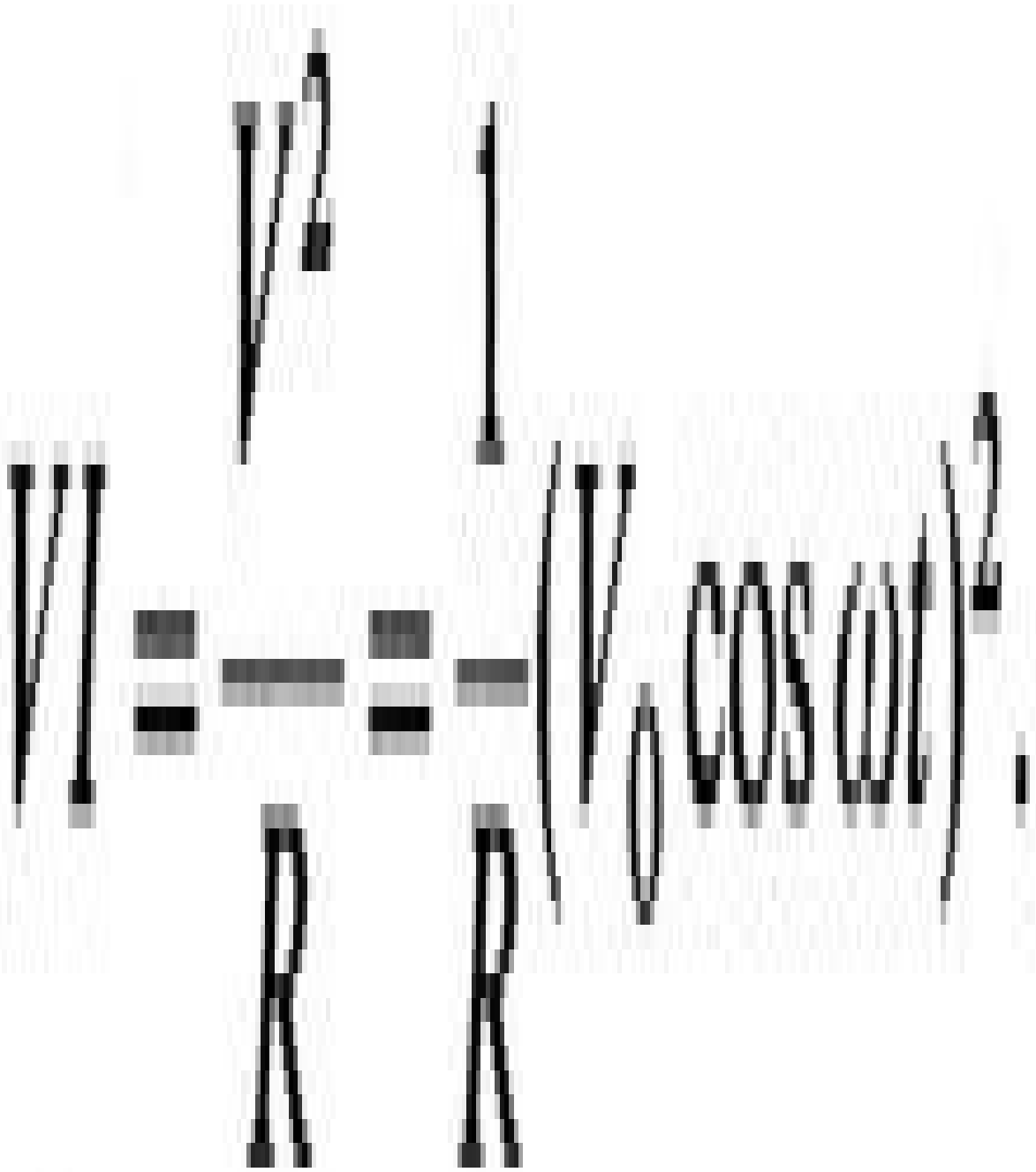
Figure 7.8. Any load made up of linear passive elements can be represented by a single impedance.


Resistance
Generator
Load
V
Z
Voltage phase lag
Reactance
Impedance

Z is the complex impedance presented to the generator by the whole network. The real part $|Z| \cos \theta$ is referred to as resistive, and the imaginary part $|Z| \sin \theta$ is called the reactance of the network. The reactance is inductive or capacitative, depending on whether the voltage leads the current or lags behind it. The phasor diagram in Figure 7.8 shows a situation in which the reactance is capacitative and the voltage lags behind the current.

7.5 POWER IN A.C. CIRCUITS

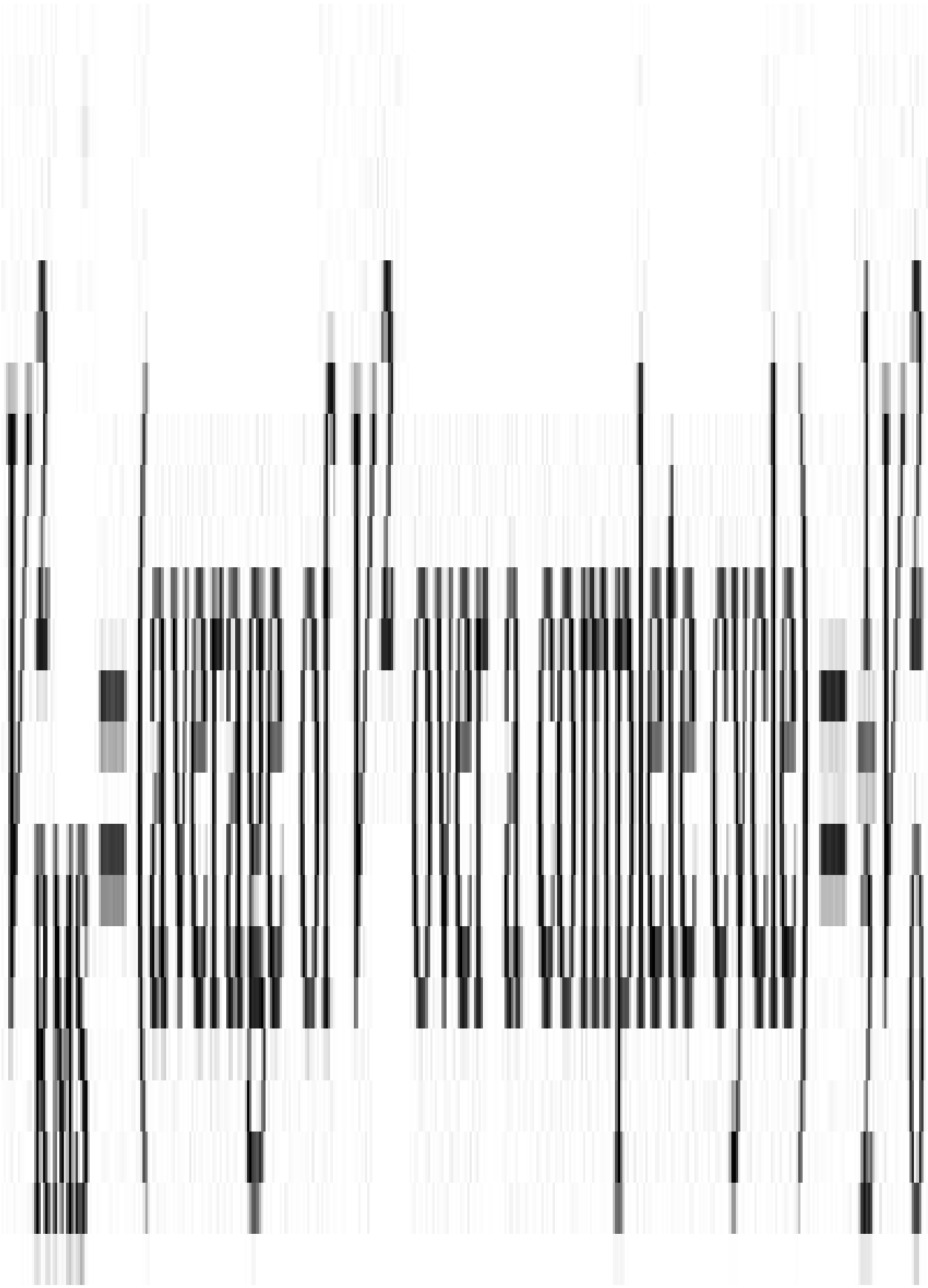
When an alternating e.m.f. drives a current through a load made up of passive elements, it must do work in order to maintain the current. In order to calculate the power demanded from an A.C. generator, all we need to know about the load is its complex impedance. Before dealing with general loads, we shall first consider the power requirements of single idealized circuit elements. The power consumption varies during the cycle, and in the resistor in Figure 7.3(a), for example, the rate of energy dissipation at time t is



The mean power W is proportional to the mean value of V^2, averaged over a complete cycle of period $2\pi/\omega$. We have

(7.14)

The mean power is half the peak power dissipated at the moment when the voltage has its maximum value V0. To avoid the occurrence of a factor in the expression for the power, it is customary to calibrate A.C. instruments in terms of root mean square (RMS) values. The root mean square voltage is defined by



or

(7.15)

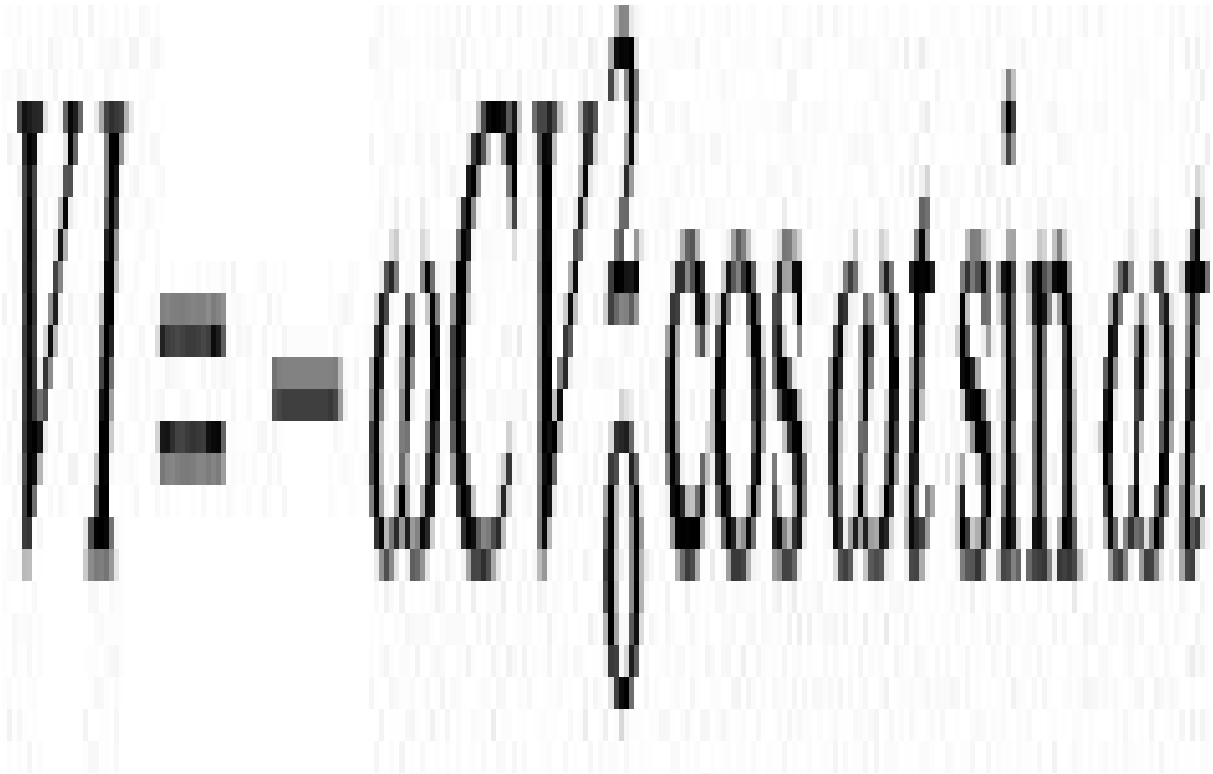
Similarly the RMS current IRMS is defined by

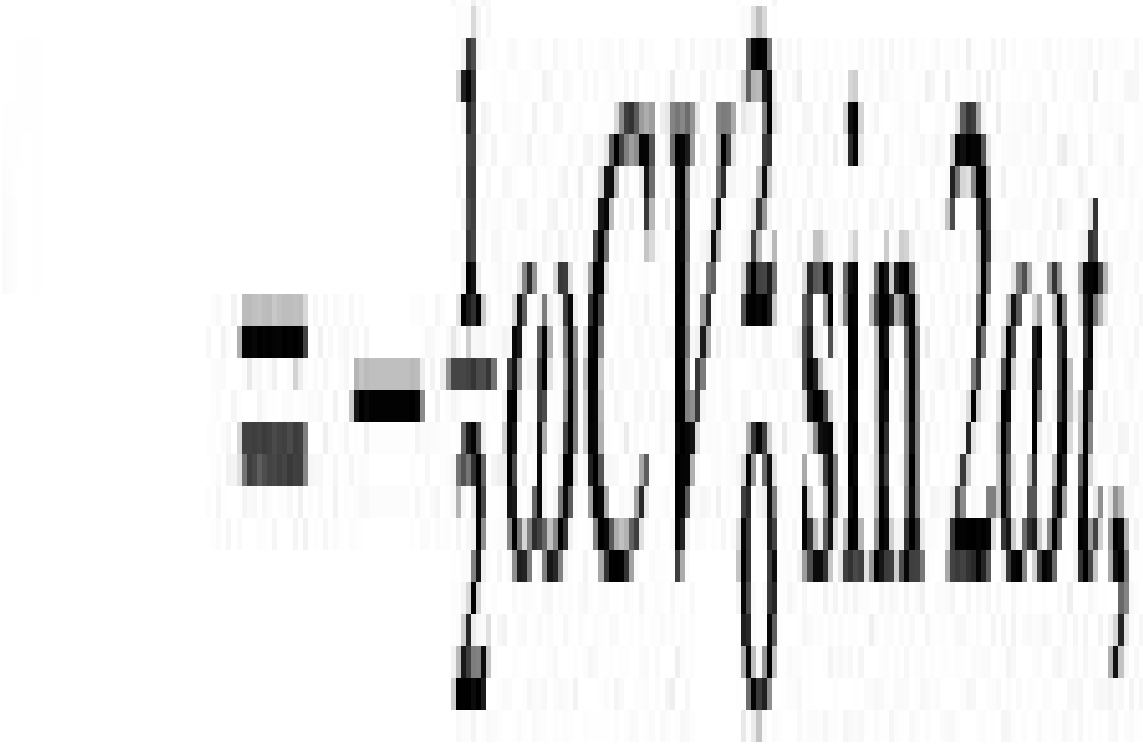
(7.16)

The RMS voltage across a resistor R carrying a current IRMS is VRMS = IRMSR. Using Equations (7.14), (7.15) and (7.16), the power dissipation can be written in the same form as for a D.C. circuit:

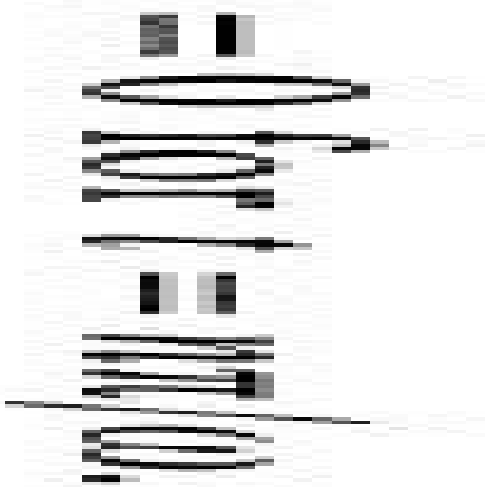
(7.17)

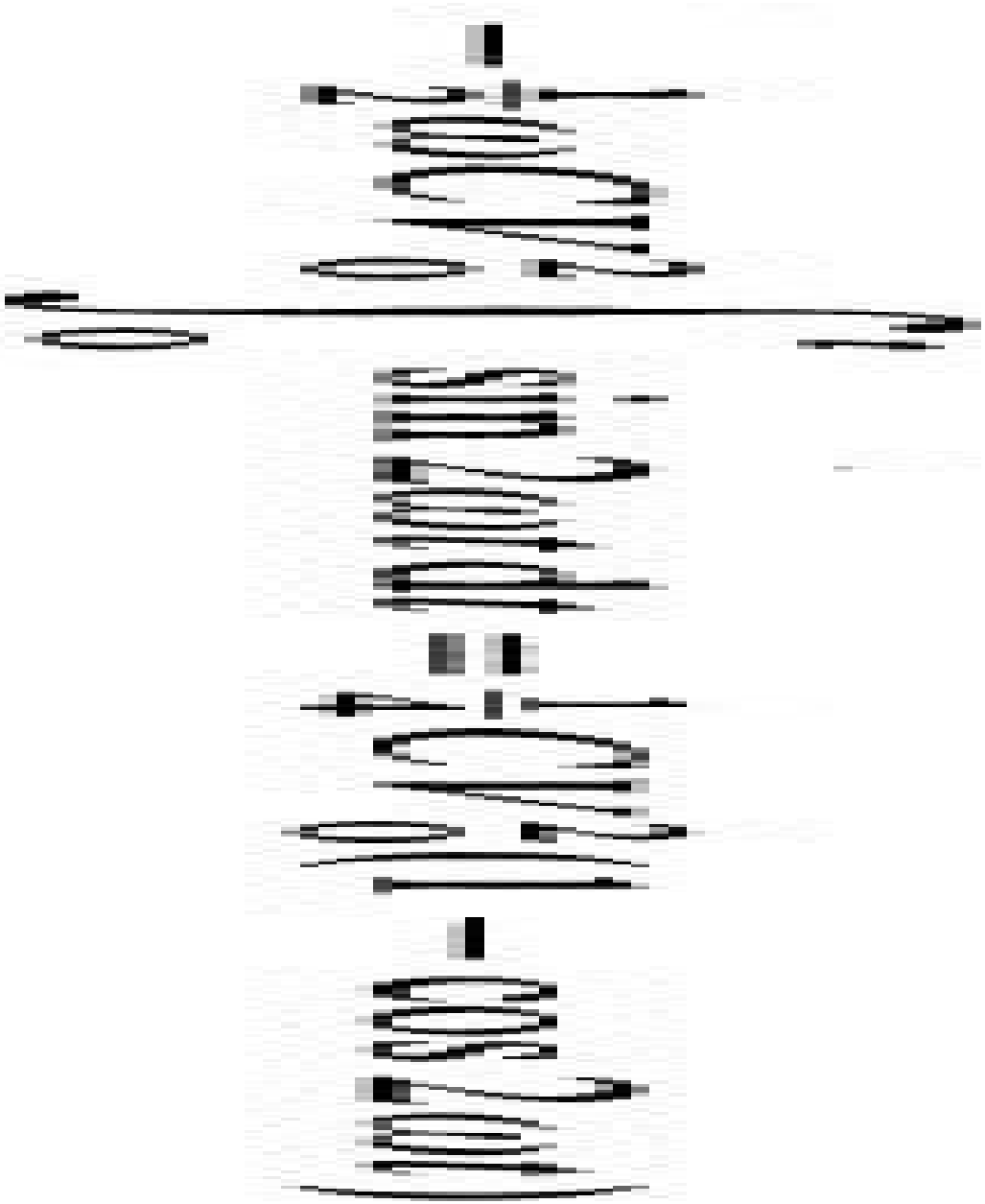
Equation (7.17) holds only for a resistor, and in an ideal capacitor or an ideal inductance no power is dissipated at all. In the capacitor circuit of Figure 7.3(b), for example, the generator delivers a voltage $V = V_0 \cos \omega t$ and a current $I = -\omega C V_0 \sin \omega t$ (Equation 7.4). The power delivered by the generator is





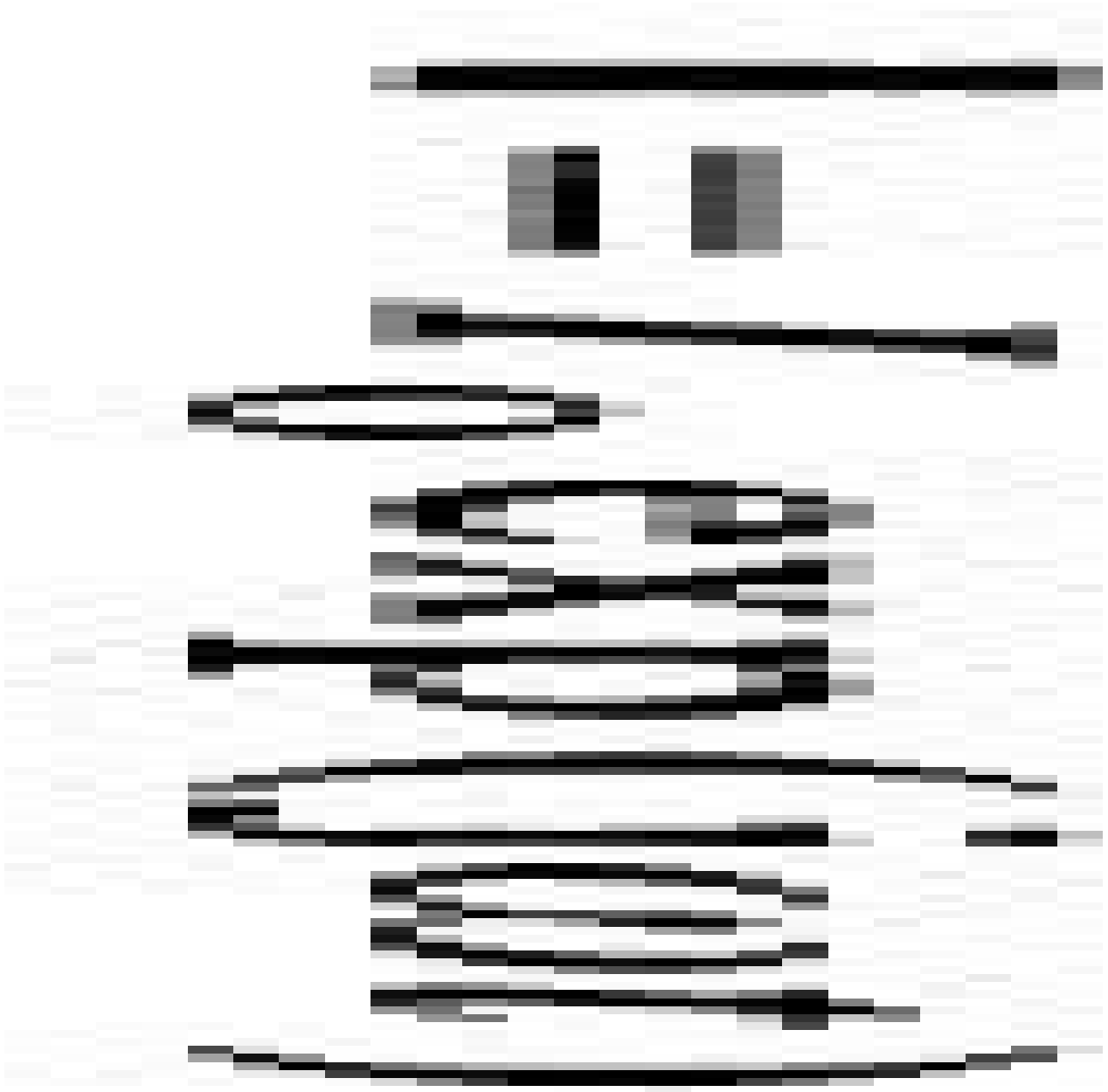
and the work done in time t is



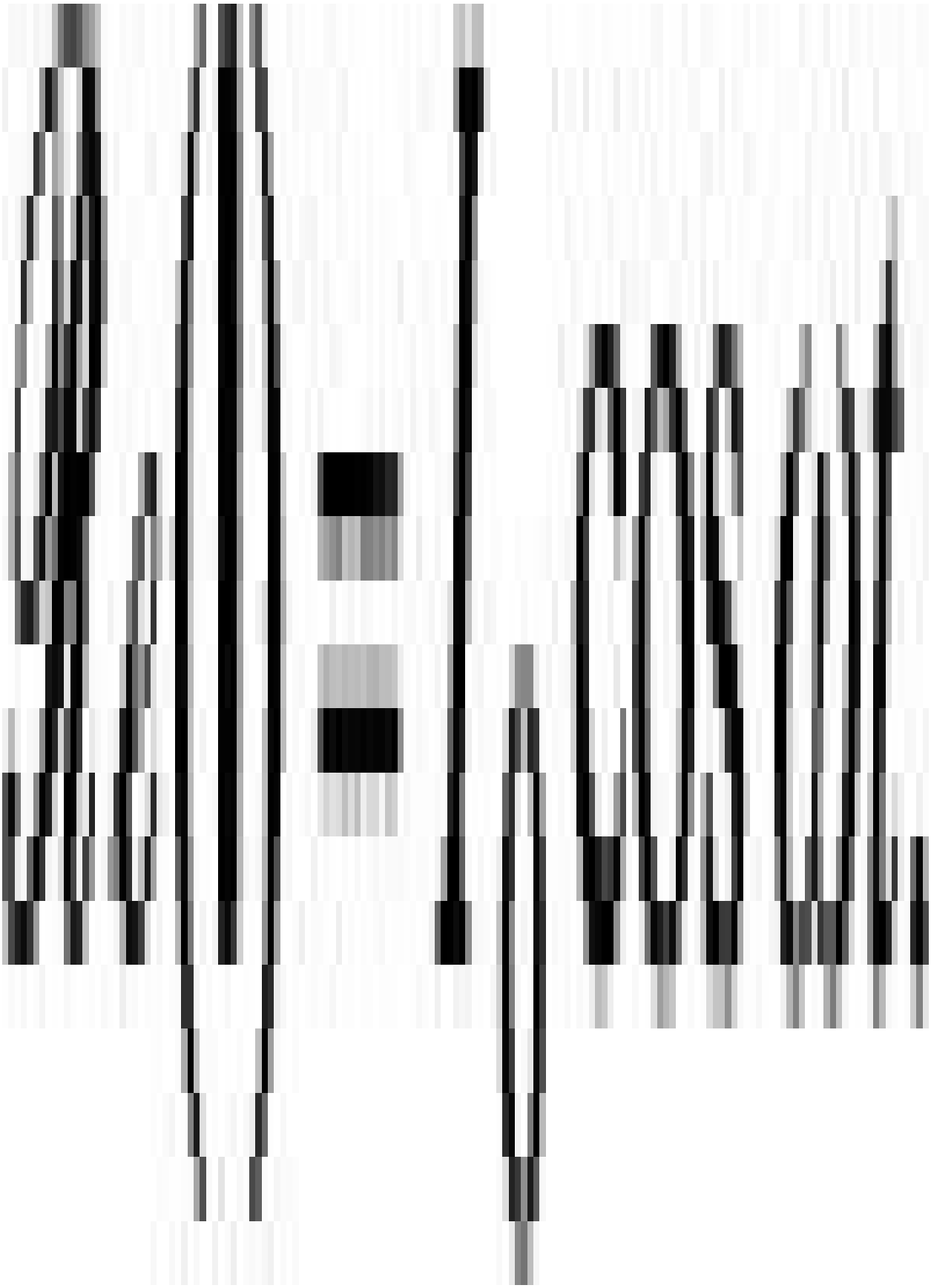


Energy surges to and fro between the generator and the capacitor, which absorbs its peak energy twice in every cycle. The power output of the generator averaged over any number of half-cycles is nevertheless zero. Similarly in the ideal inductance shown in Figure 7.3(c) the stored magnetic energy builds up to twice in every cycle, and this energy is returned to the generator as the current falls to zero.

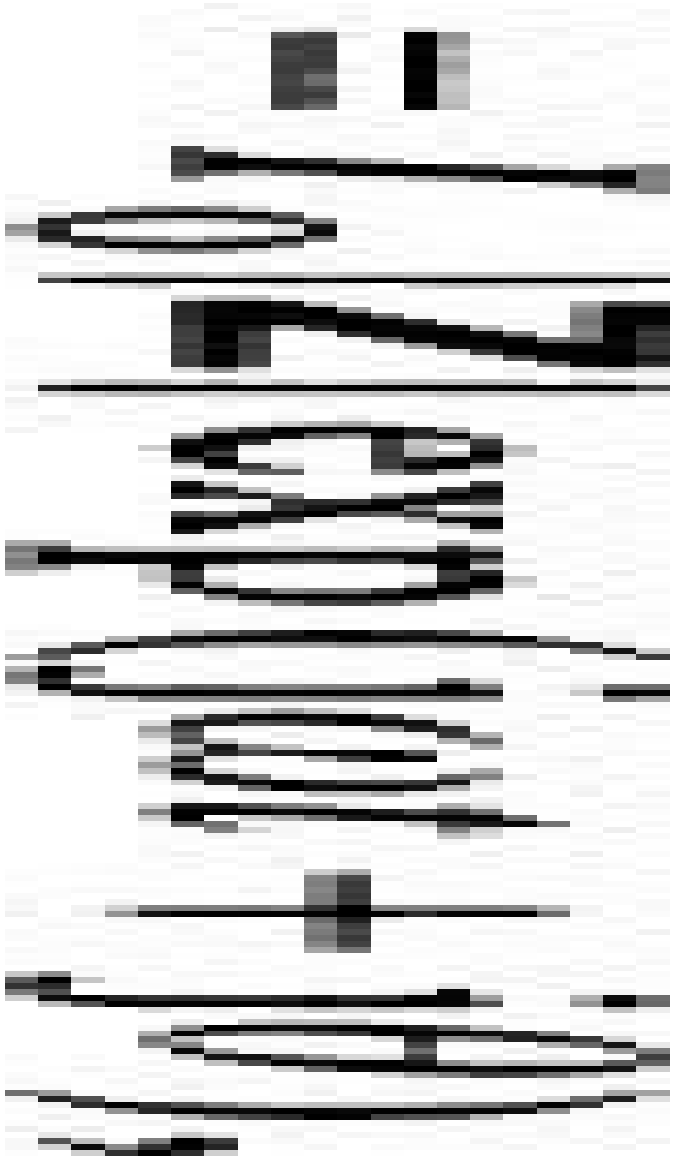
Now we can calculate the power required to drive current through any load made up of passive circuit elements. We shall discover that only the resistive part of the load absorbs power, and that the reactance simply stores energy, and later returns it. Consider the load circuit in Figure 7.8, with a complex impedance Z. We must be rather careful in using complex numbers when working out power dissipation. Although it is permissible to add complex impedances in series, as we did for the equivalent circuit of the coil with a finite resistance, remember that only the real part of the quantities appearing in complex equations has any physical meaning. Before multiplying voltage and current to find the power, it is necessary to throw away the imaginary parts to avoid having a spurious contribution to the product. Again choosing the time-zero when the current is maximum,

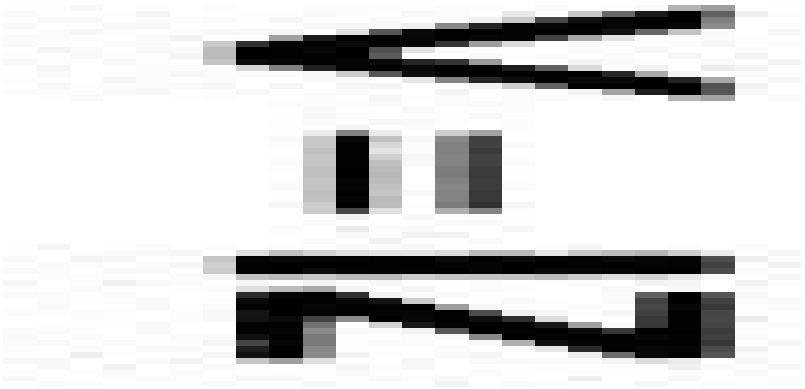


and

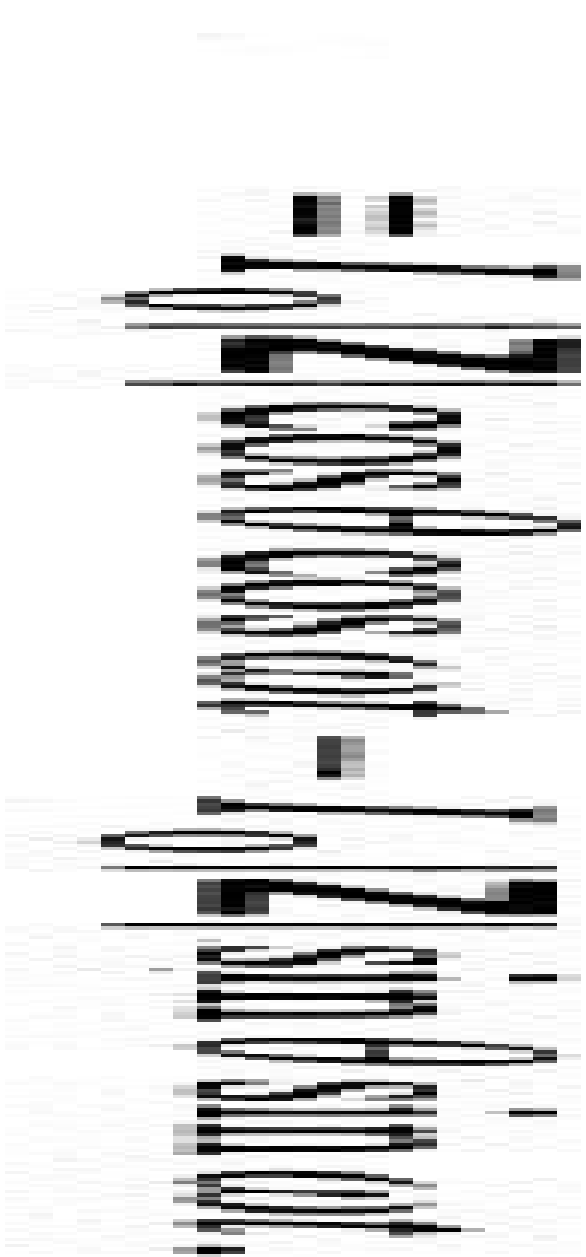


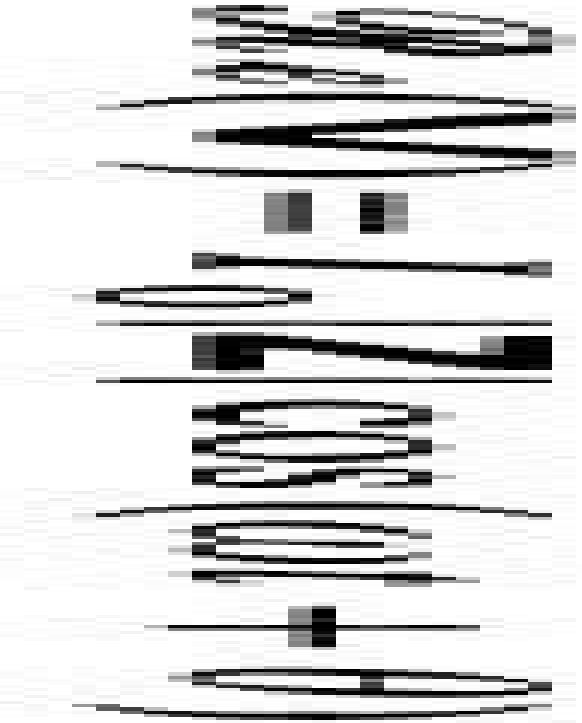
The voltage is



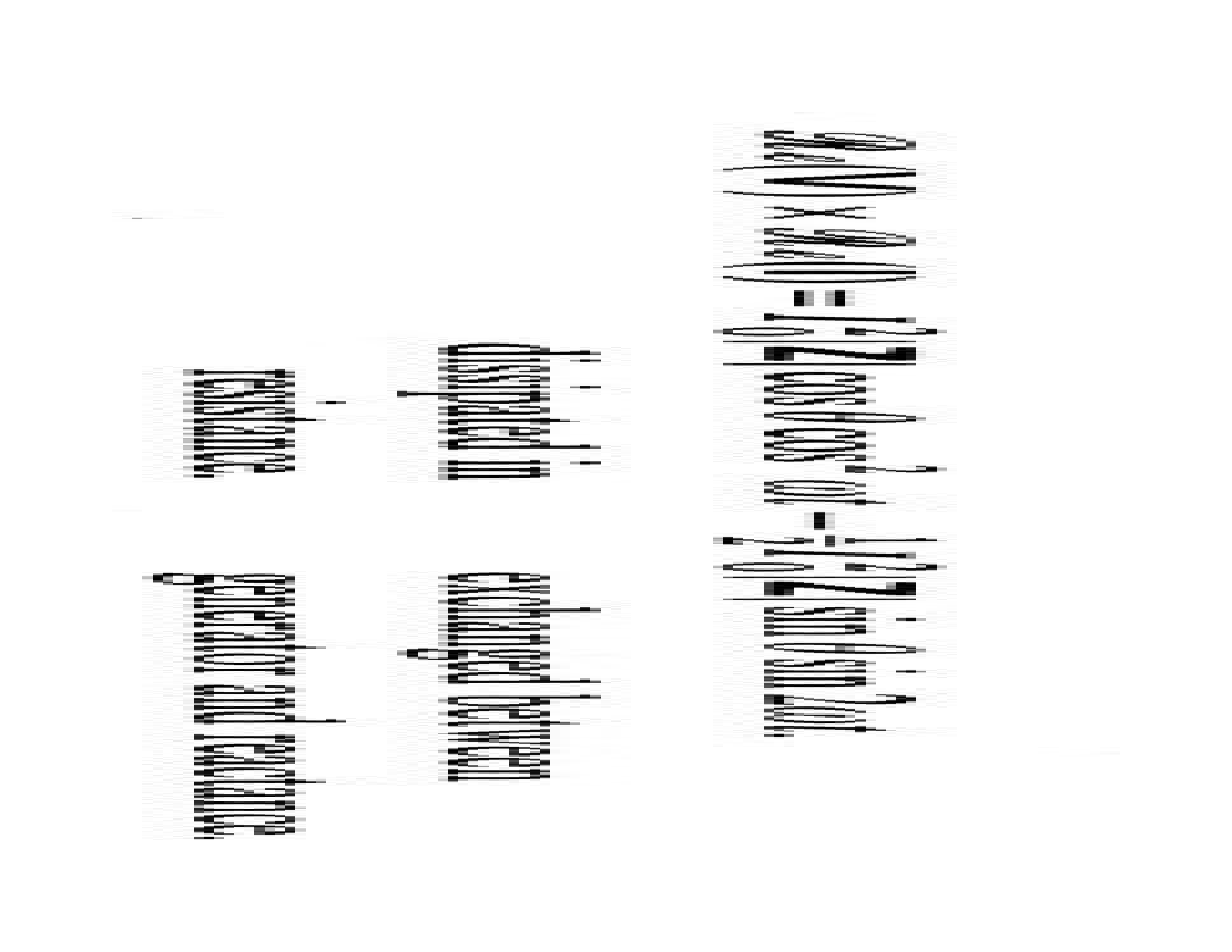


and





The power drawn from the generator at time t is

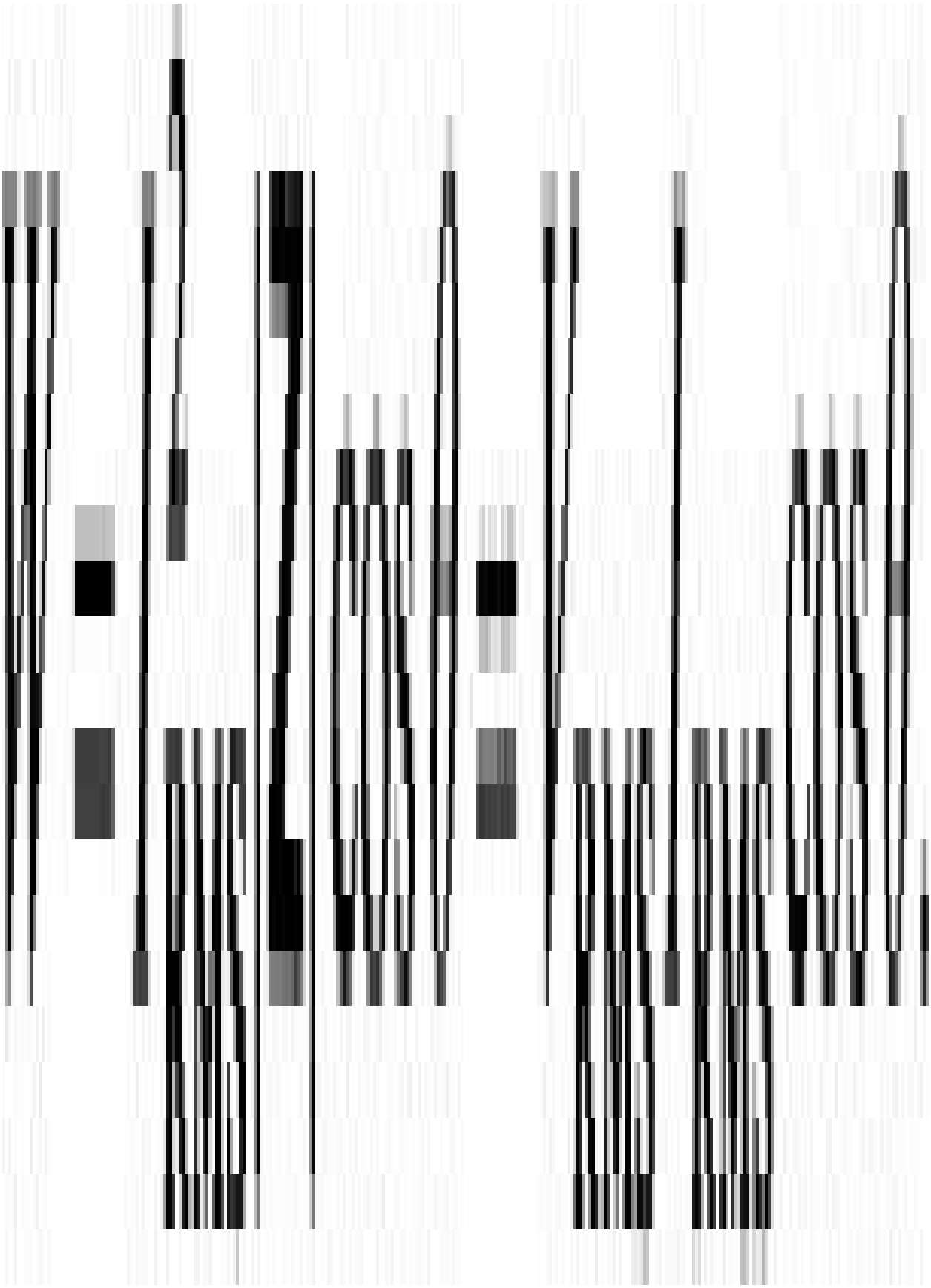


Averaged over a complete cycle, the power dissipated in the load is

(7.18)

or in terms of RMS values

(7.19)

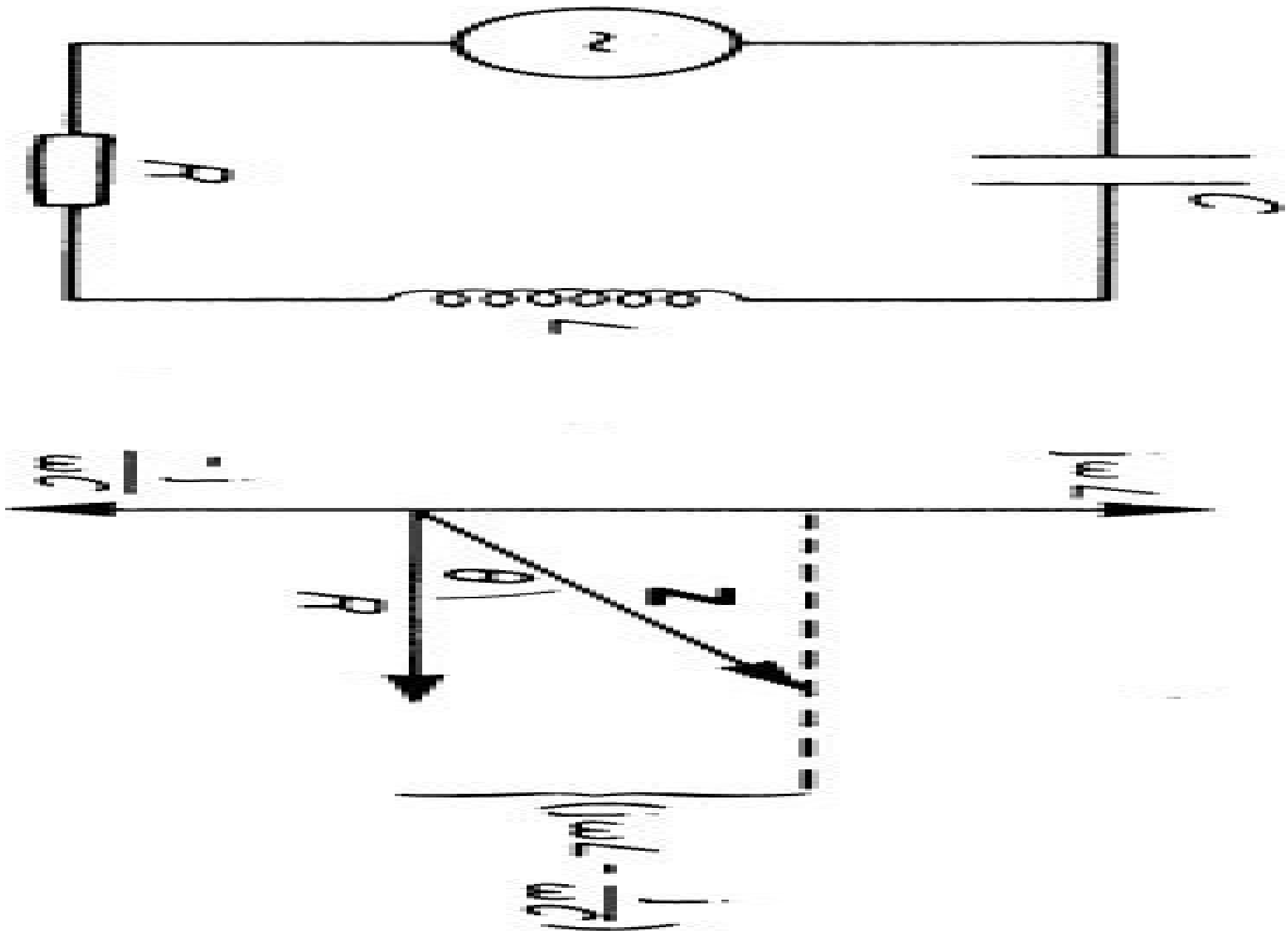


VRMS and IRMS are the values registered by an A.C. voltmeter and an A.C. ammeter measuring the output of the generator. For a purely resistive load the power is given directly by their product, but for a reactive load, the power is reduced by the power factor cos θ. In a typical coil with $\theta = 86°$ at the main supply frequency, the power dissipation is less by a factor of ten than in a non-reactive component whose impedance has the same magnitude.

7.6 RESONANCE

A capacitor presents a large impedance to low frequency alternating current, but has a small impedance at high frequency. The opposite is true for an inductive coil which resists changes in current, and generates an e.m.f. proportional to the frequency. What happens when an inductive coil and a capacitor are connected together in series across an A.C. generator, as in the circuit diagram of Figure 7.9? The impedance is now large both at high frequencies (because of the coil) and at low frequencies (because of the capacitor). A minimum value of the impedance occurs at some intermediate frequency called the resonant frequency of the circuit. We shall find that if the resistance of the coil windings is sufficiently low, then the impedance increases rapidly on either side of the minimum as the frequency moves away from the resonant frequency. A coil and a capacitor connected in series can select signals near their resonant frequency while rejecting signals of other frequencies for which the impedance is high. The circuit is said to be ‘tuned’ to the resonant frequency: tuned circuits are used in radio receivers to discriminate between stations broadcasting on different wavelengths. To investigate the sharpness of tuning (i.e. the width of the frequency band over which signals are accepted), we must find the resonant frequency and work out the variation of impedance near this frequency.

Figure 7.9. The series resonant circuit.



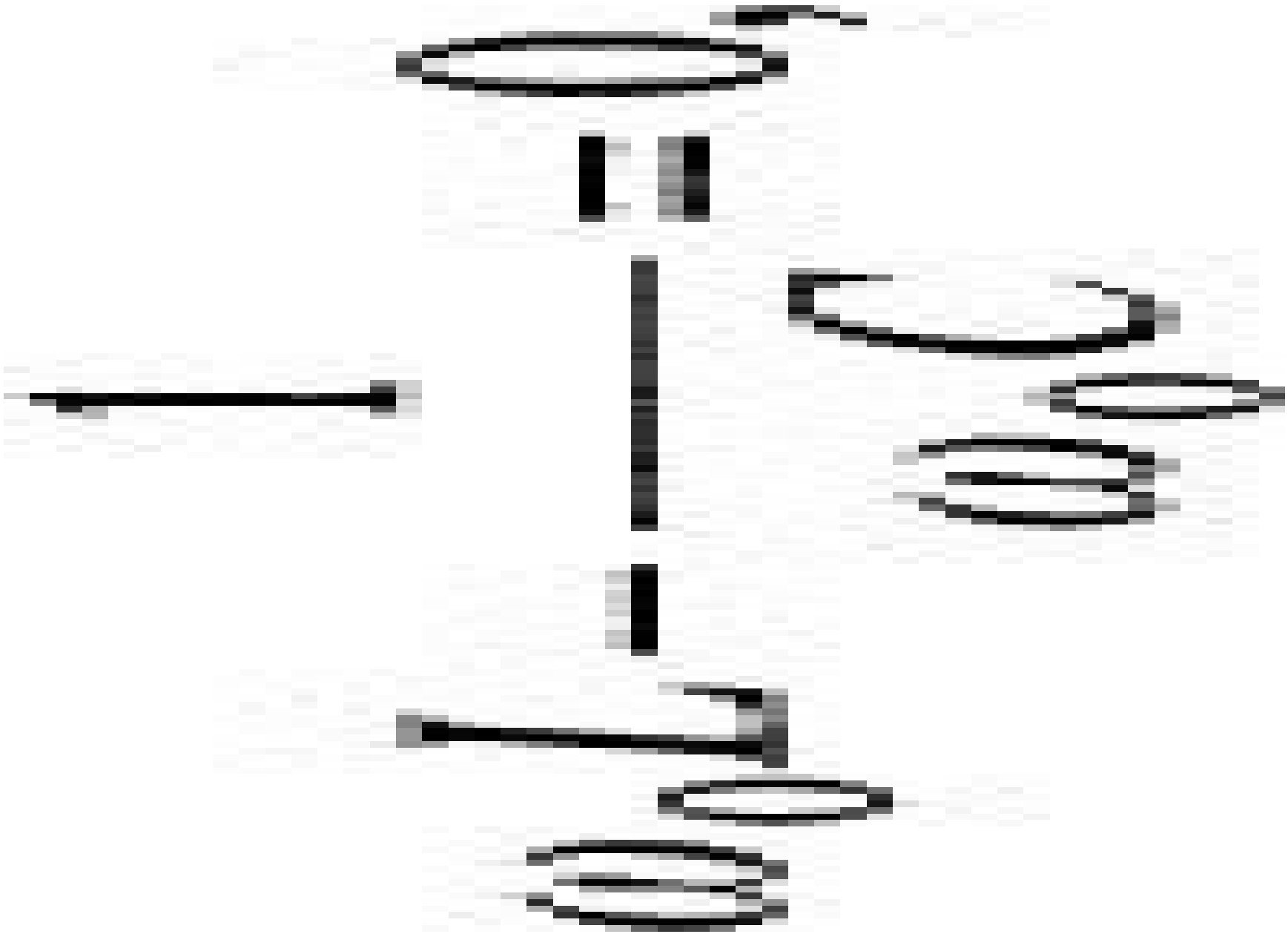
The impedance diagram of the resonant circuit is shown in Figure 7.9. The complex impedance Z is the sum of the impedances of the components L, C and R separately. At an angular frequency ω

(7.20)

The magnitude of the impedance is

(7.21)

which has its minimum value at the resonant angular frequency ω0 given by

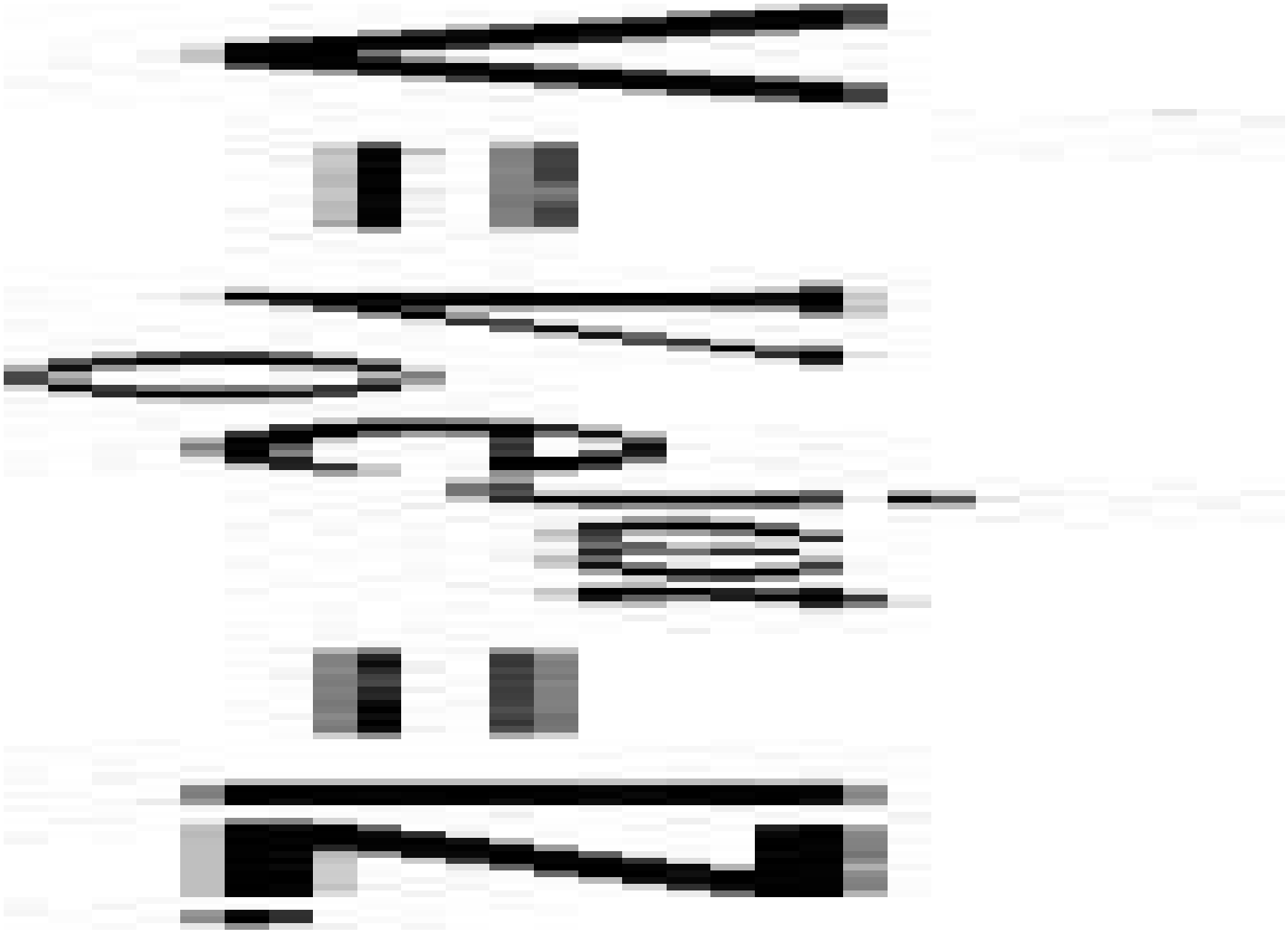


or

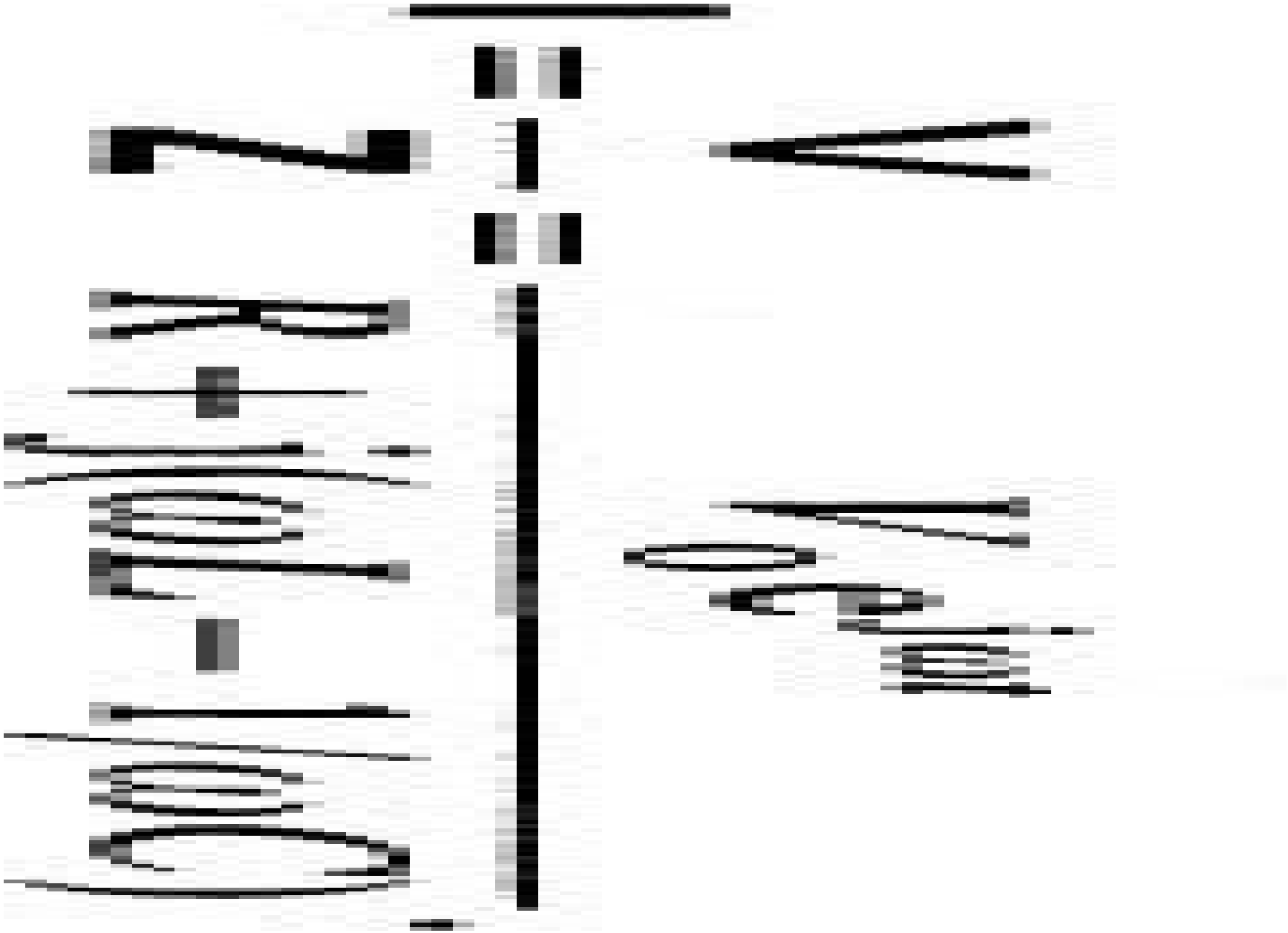
(7.22)

Resonance occurs at the frequency at which the inductive and capacitative impedances are equal and opposite, so that they exactly cancel one another. The impedance at resonance is real, having the value R, as though a resistor R were the only component in the circuit.

Let us assume that the generator voltage has a constant amplitude V0, independent of frequency. Now we can find the current I from the relation

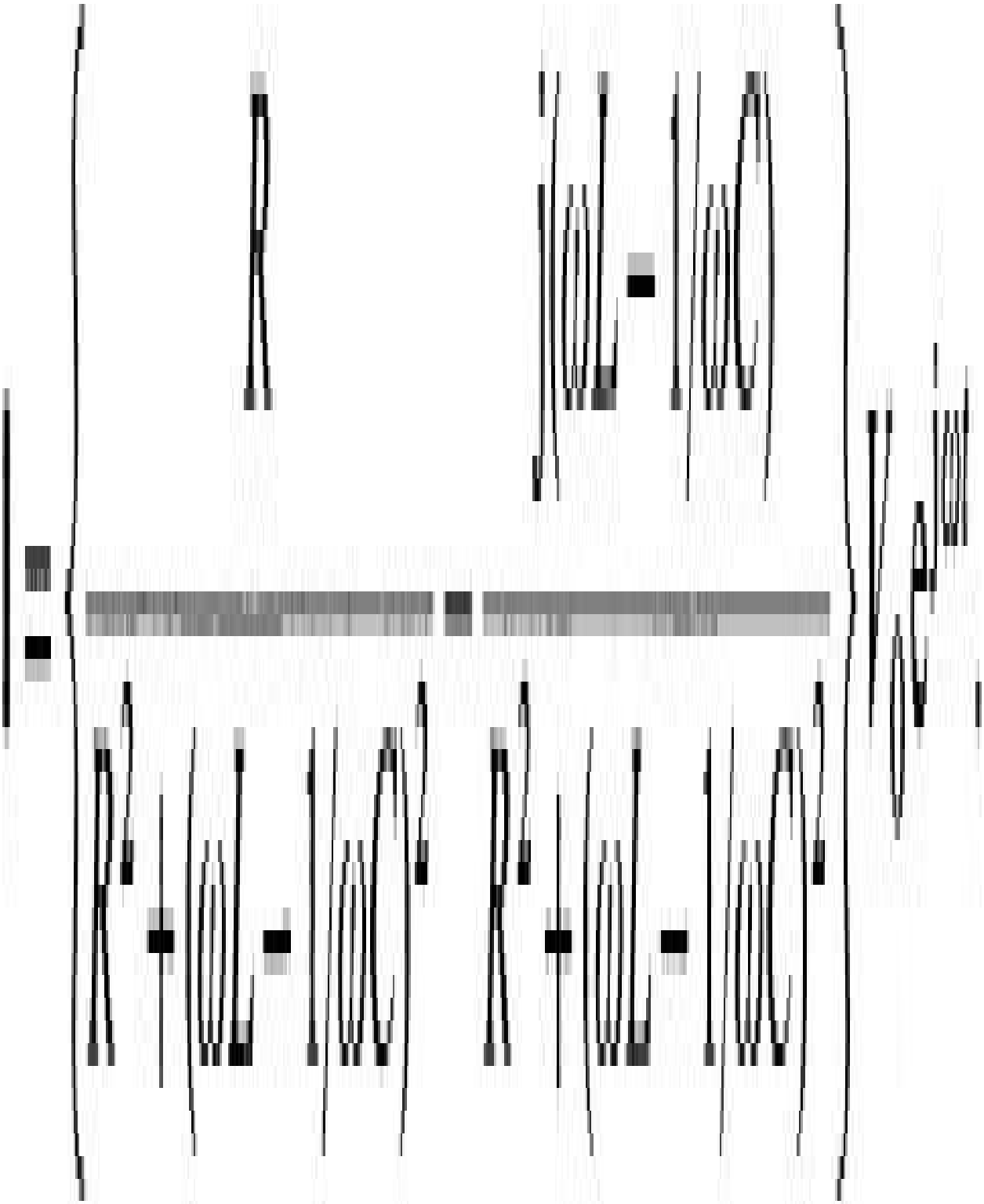


Hence

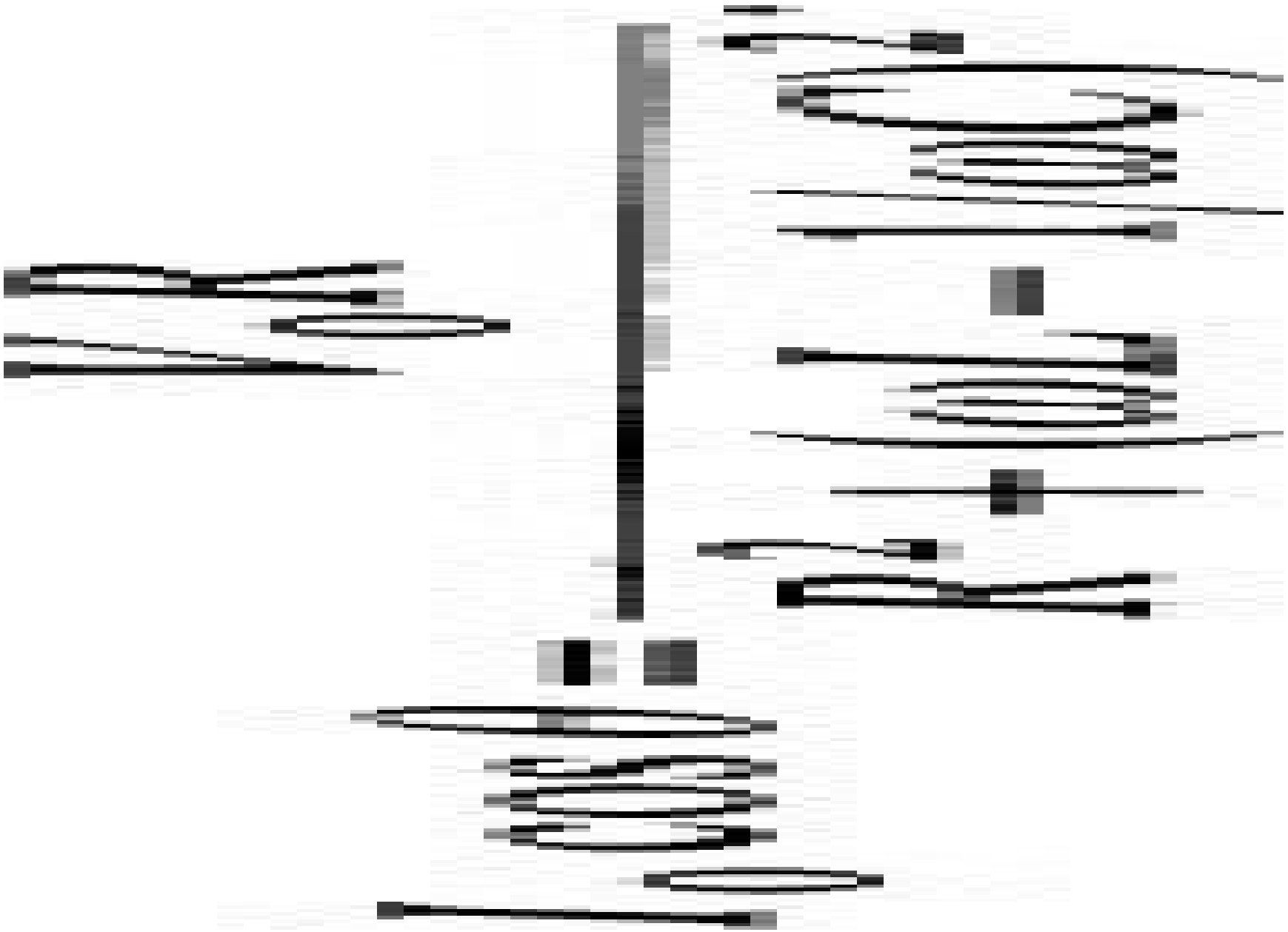


Multiplying top and bottom by $\{R - j(\omega L - 1/\omega C)\}$,

(7.23)

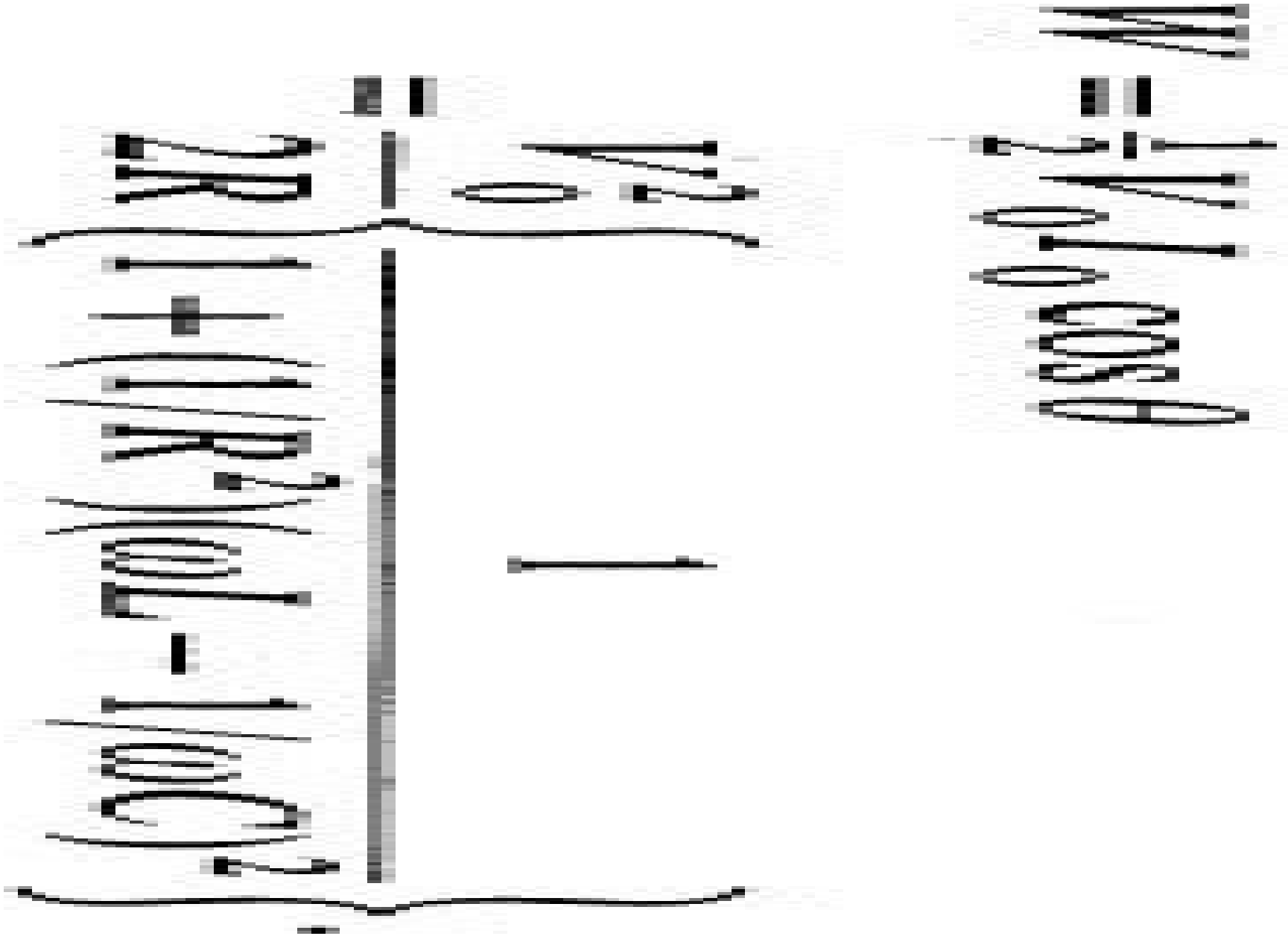


In the previous section we showed that there is a net demand for power from the generator only from that component of the current which is in phase with the voltage, namely the first term in Equation (7.23). If the current has amplitude I_0 and lags behind the voltage by a phase angle θ, then the component in phase with the voltage is

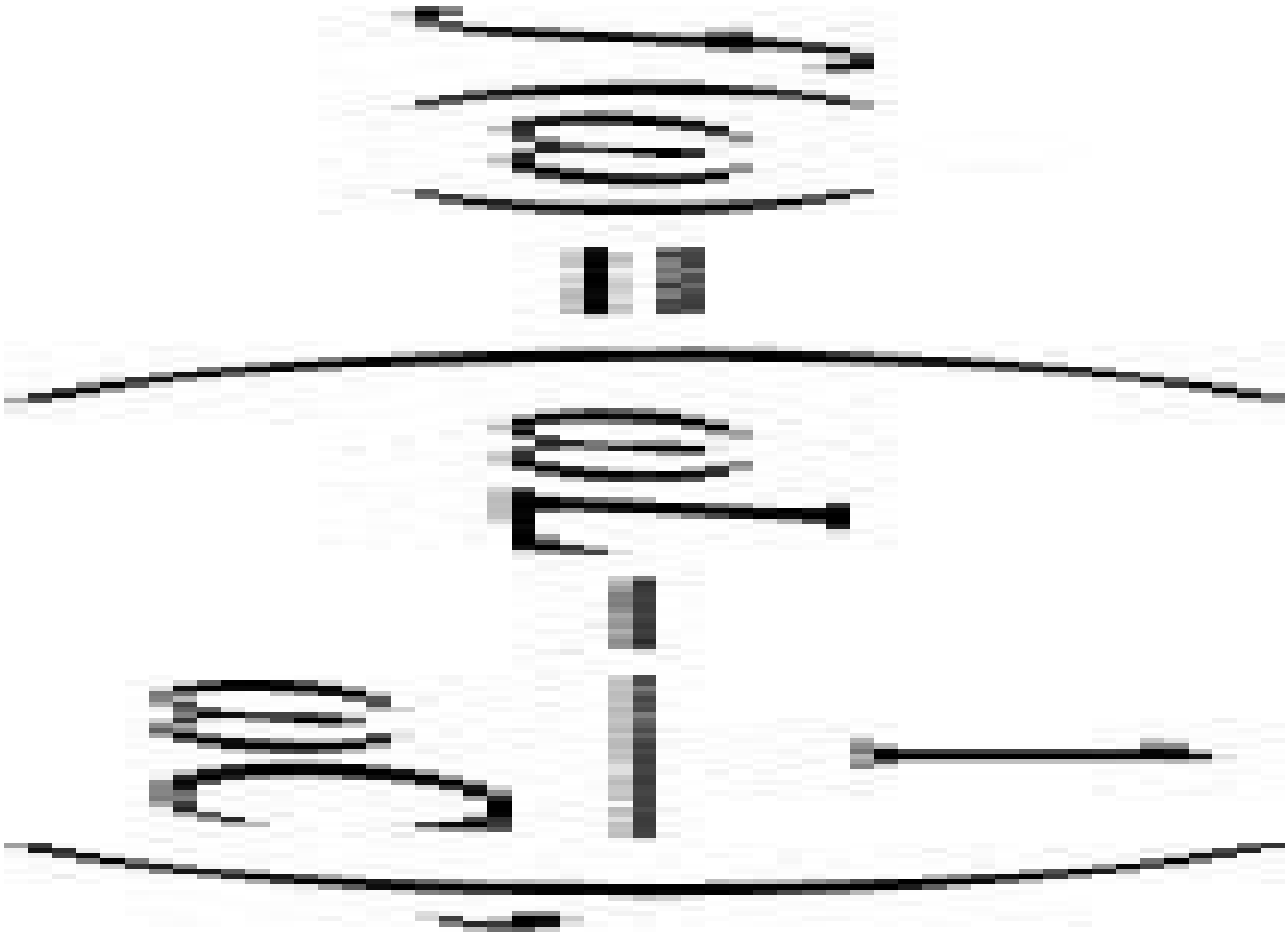


The power W, averaged over a complete cycle, is given by Equation (7.18):

(7.24)



Away from resonance the power falls off rapidly, as shown in Figure 7.10. Restricting attention to angular frequencies near to ω_0, we can make a Taylor expansion and find an approximate expression for W by neglecting high order terms. Writing



we have

$$f(\omega) \simeq f(\omega_0) + (\omega - \omega_0) \left. \frac{\partial f}{\partial \omega} \right|_{\omega_0}$$

$$= (\omega - \omega_0) \left(L + \frac{1}{\omega_0^2 C} \right)$$

$$= 2(\omega - \omega_0) L,$$

since $f(\omega 0) = 0$ and . Substituting in Equation (7.24), the average power delivered by the generator, or equivalently the average power absorbed by the circuit, is approximately

(7.25)

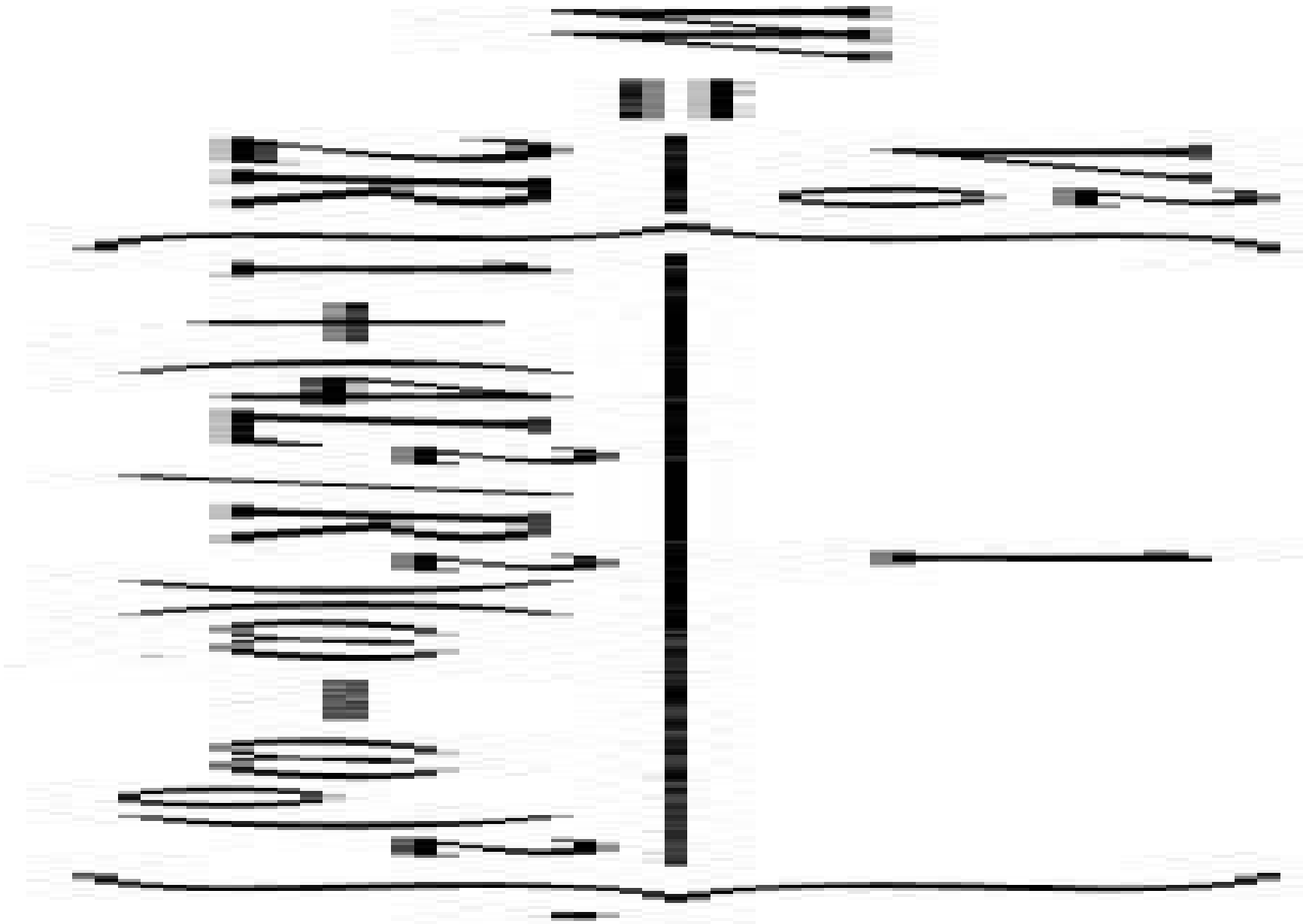
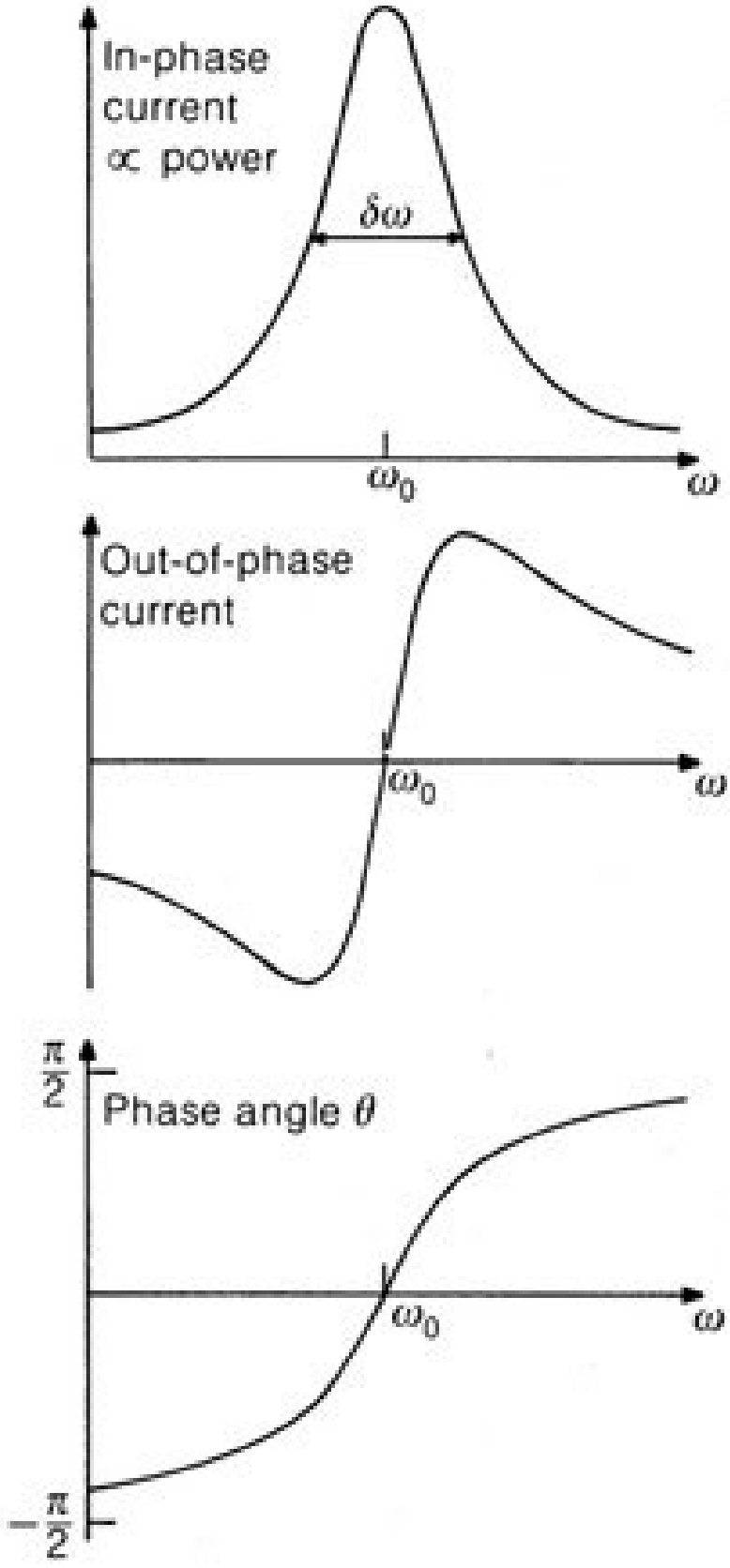
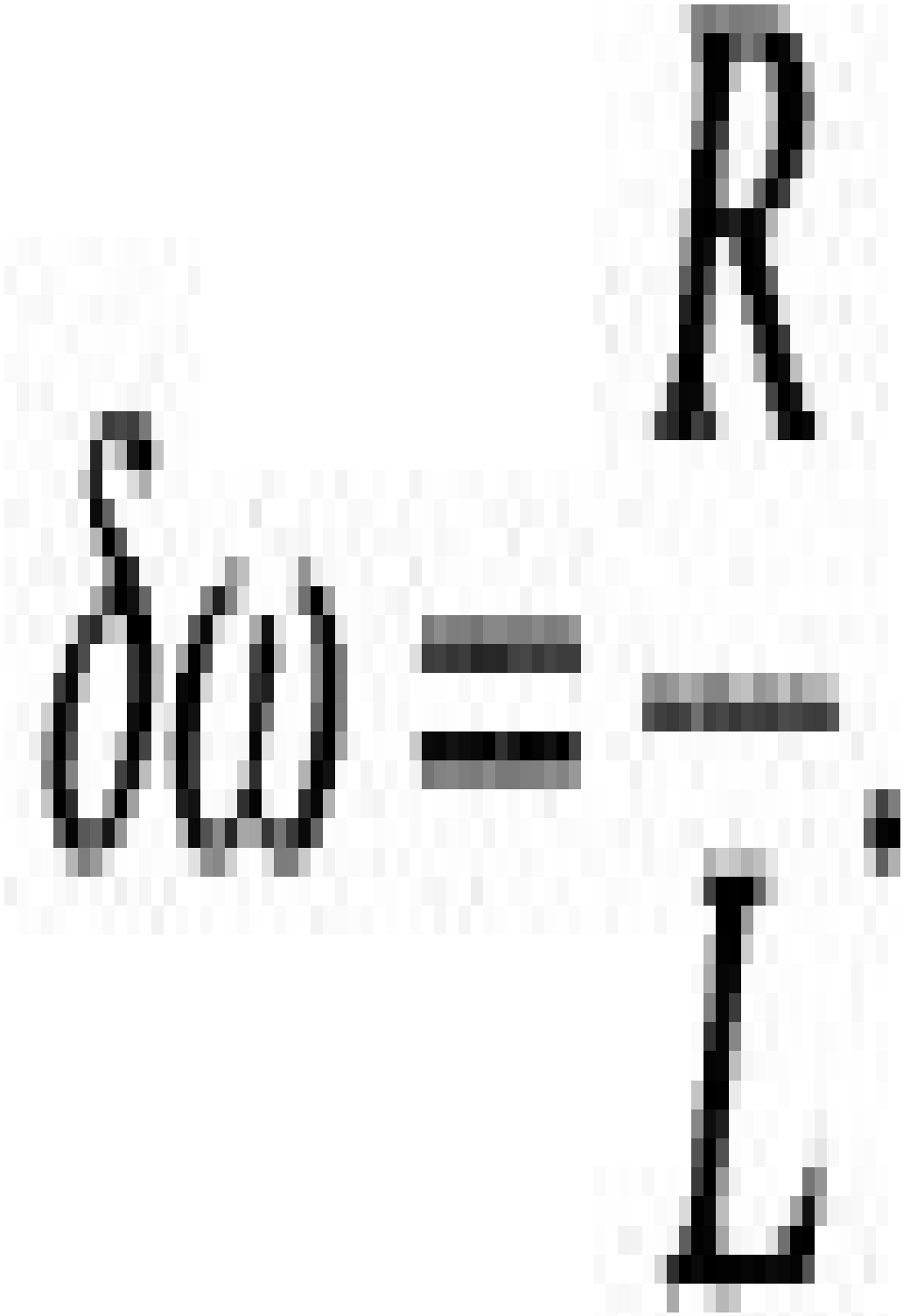


Figure 7.10. Currents and phase angles near resonance.


In-phase
current
∝ power
δω
ω0
ω
Out-of-phase
current
ω0
ω
π/2
Phase angle θ
ω0
ω
−π/2

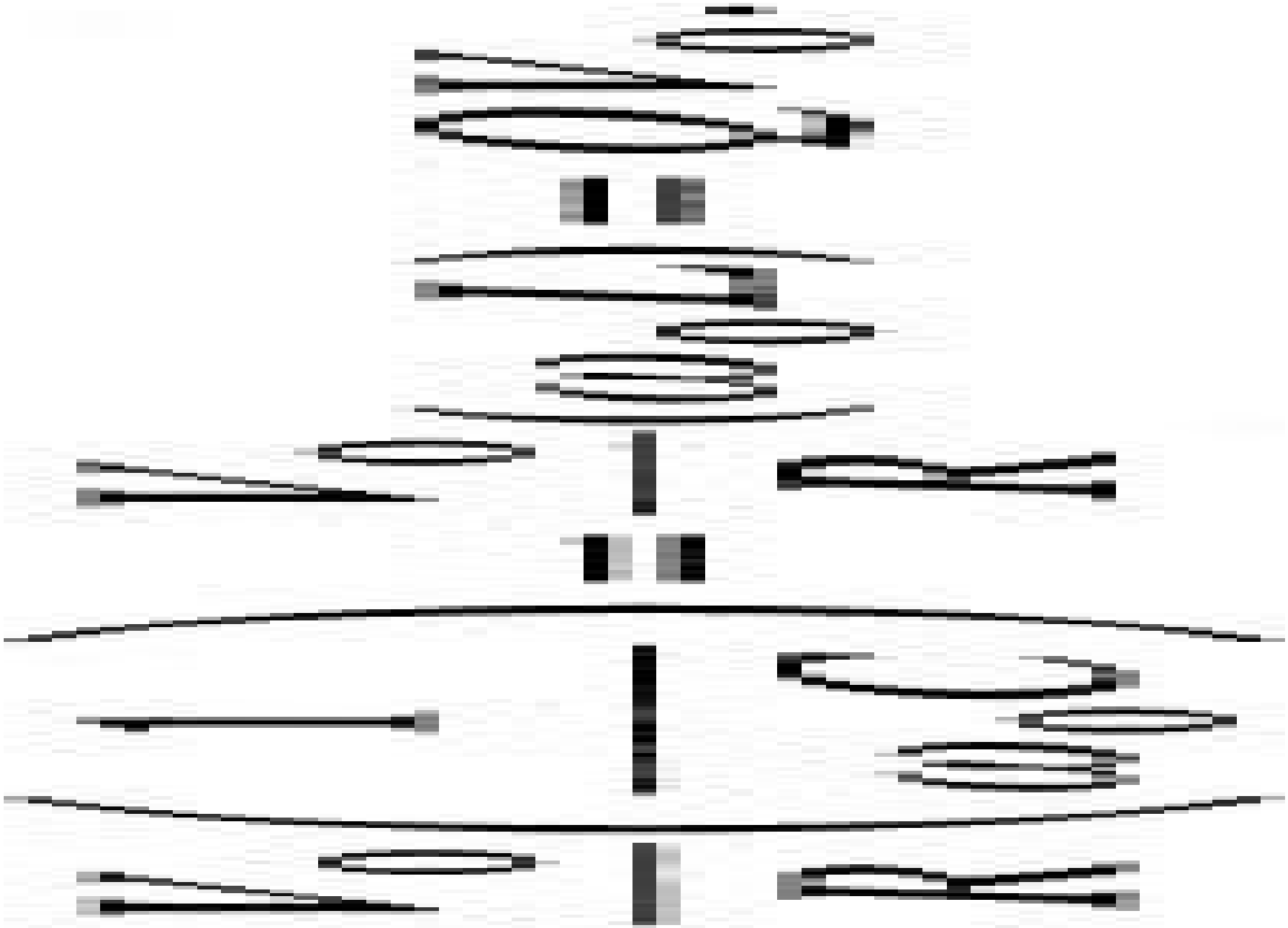
The power falls to half its maximum value when $(\omega - \omega 0) = \pm R/2L$, and the angular frequency band over which the power absorption is greater than half maximum is



The angular frequency band $\delta\omega$ for the series resonant circuit is usually expressed as a fraction of the resonant angular frequency ω_0;

(7.26)

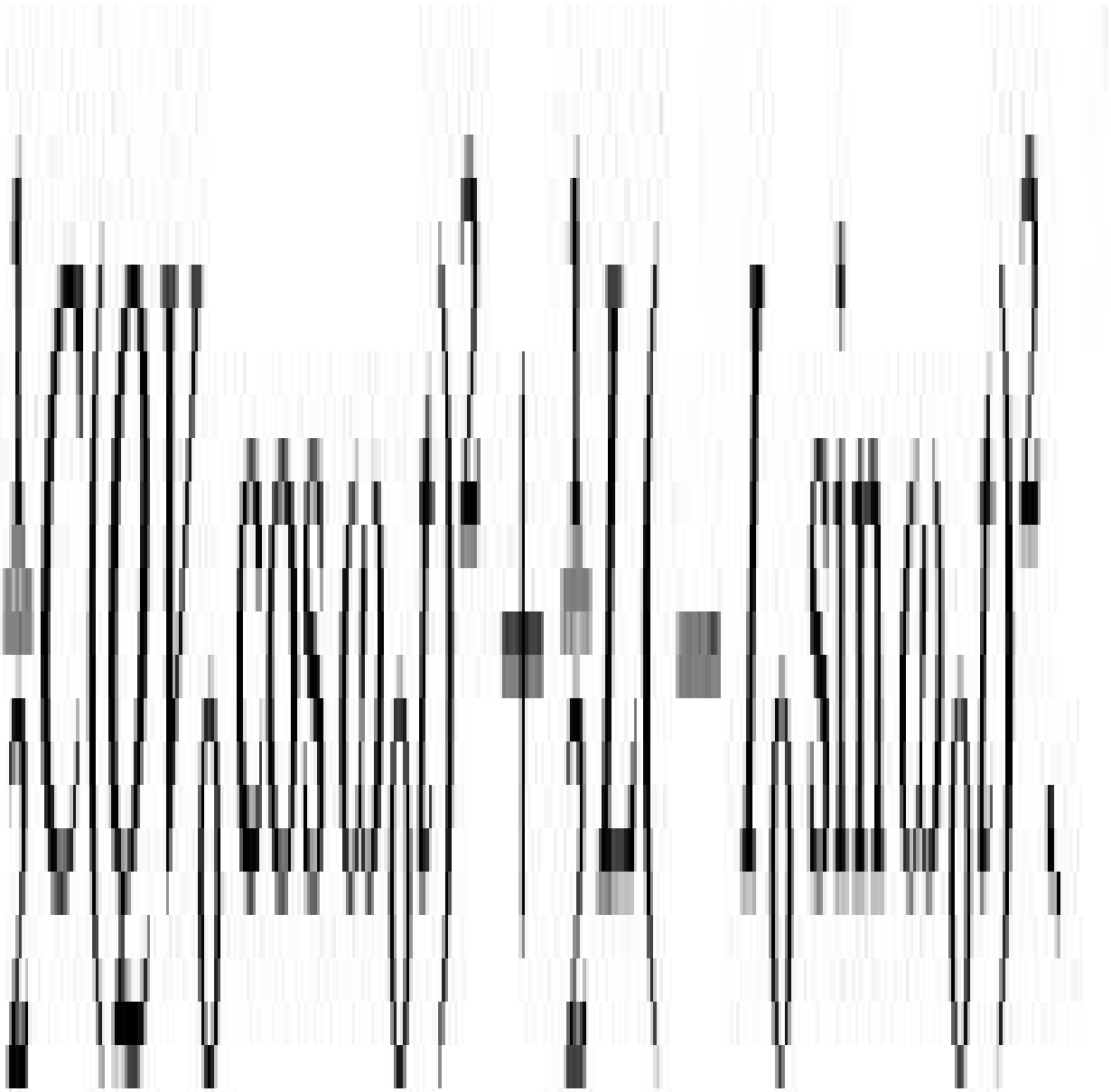
The symbol Q stands for ‘quality factor’ (almost invariably left in its abbreviated form as ‘Q’). Q is the ratio of the inductive impedance of the coil to the resistance of its windings, and we see from Equation (7.26) that the larger the value of Q, the narrower is the range of angular frequency over which the circuit absorbs much power. Near resonance, voltages much larger than V_0 appear across the coil and the capacitor. At the angular frequency ω_0, the current amplitude in the circuit is V_0/R, and it follows that the amplitude of the voltage across both capacitor and inductance is



A circuit with a high value of Q can thus be tuned to select an incoming signal of a particular frequency. Either C or L may be altered until resonance occurs at this frequency. There is then a large voltage across both the capacitor and the coil, while other frequencies are attenuated. In practice one cannot construct high-Q circuits at low frequency because, even when using iron-cored coils, the resistance of the windings is such that the ratio L/R is usually about 0.05 henrys/ohm, and thus

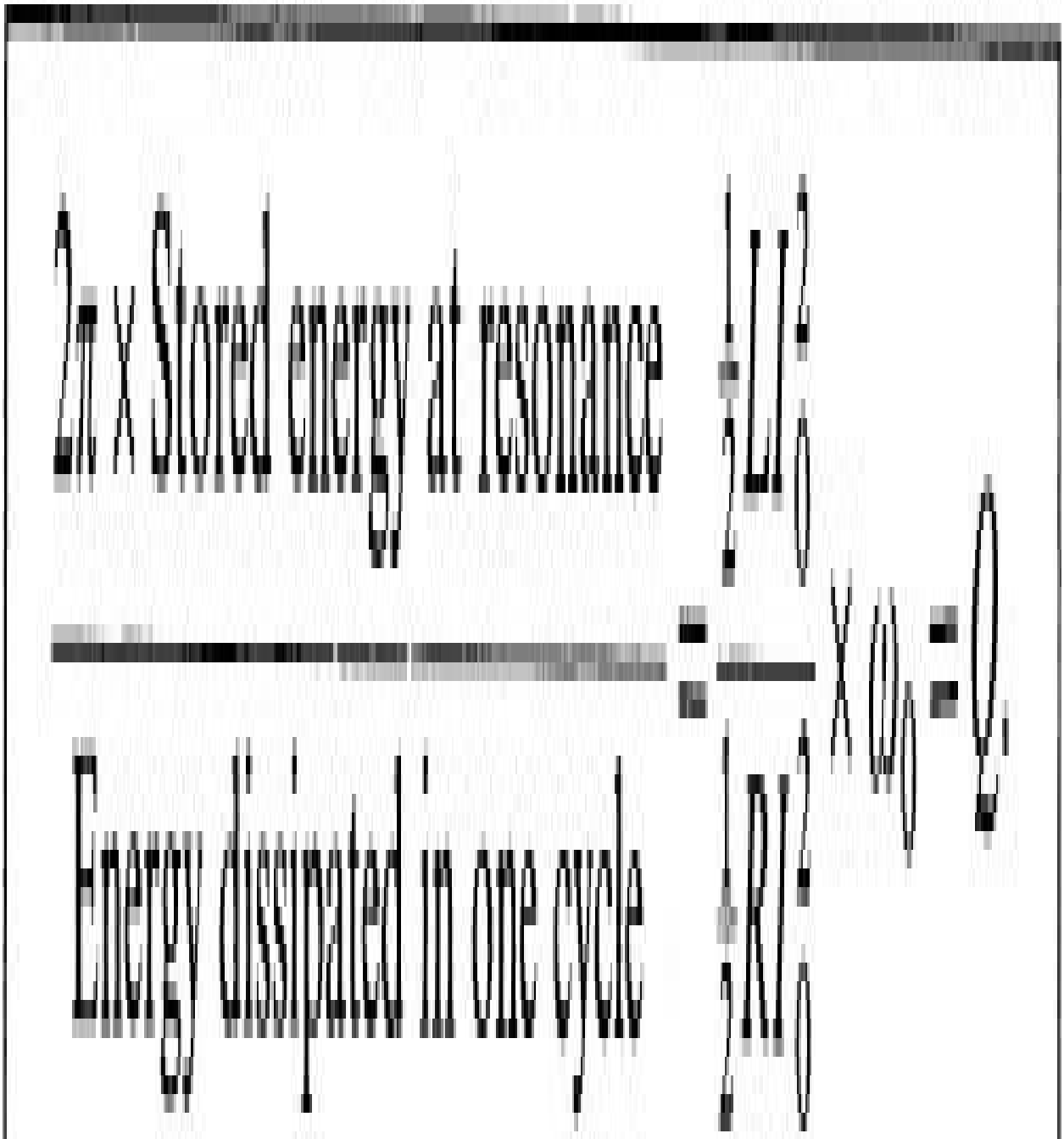
For example Q 15 at the main supply frequency. However, at the frequencies used in radio transmission, high Q-values and sharp tuning are easily obtained.

Energy is needed to build up the voltage across the capacitor, and to drive current against the inductive impedance of the coil. We shall discuss in the next section how the energy builds up when the generator is first switched on; once the alternating current is established, however, the energy is stored by the circuit and surges to and fro between capacitor and coil. To find the amount of stored energy at resonance, we drop the complex notation, writing the voltage across the capacitor as ($QV_0 \cos \omega_0 t$), and the current, which is out of phase by $\pi/2$, as $I_0 \cos(\omega_0 t + \pi/2) = -I_0 \sin \omega_0 t$. The stored energy is the sum of the electrostatic energy and the magnetic energy, i.e.



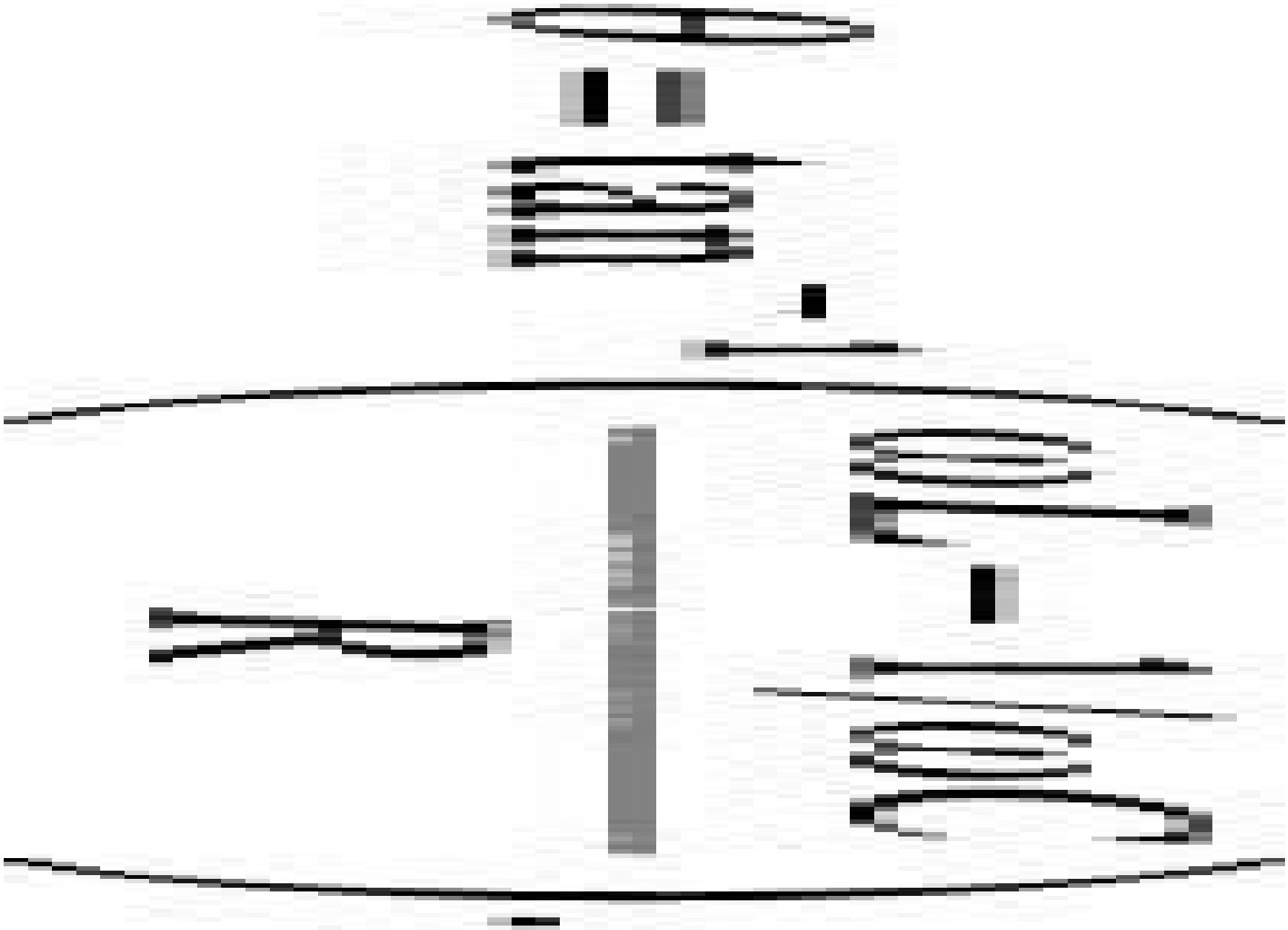
After replacing V0 by I0R, a little manipulation shows that the electrostatic energy is . The total energy stored is therefore . Comparing the stored energy with the energy dissipated in the resistance during the time $1/\omega_0$ taken for the phase to advance by one radian, which is $1/2\pi$ times the period of a complete cycle, we find

(7.27)


2π x Stored energy at resonance
Energy dissipated in one cycle

Q is equally well defined in terms of energy by this relation as by Equation (7.26).

Away from resonance, the peak energies stored by the capacitor and the coil are not equal, and the circuit must exchange the difference with the generator. This energy is supplied by the component of the current which is $\pi/2$ out of phase with the voltage, namely the second term in Equation (7.23). The magnitude of the out-of-phase current, which is plotted in Figure 7.10, changes sign at resonance, as the reactance changes from being capacitative to being inductive. Also shown in the figure is the phase angle by which the voltage leads the current. From the diagram in Figure 7.9 we can see that the phase angle is



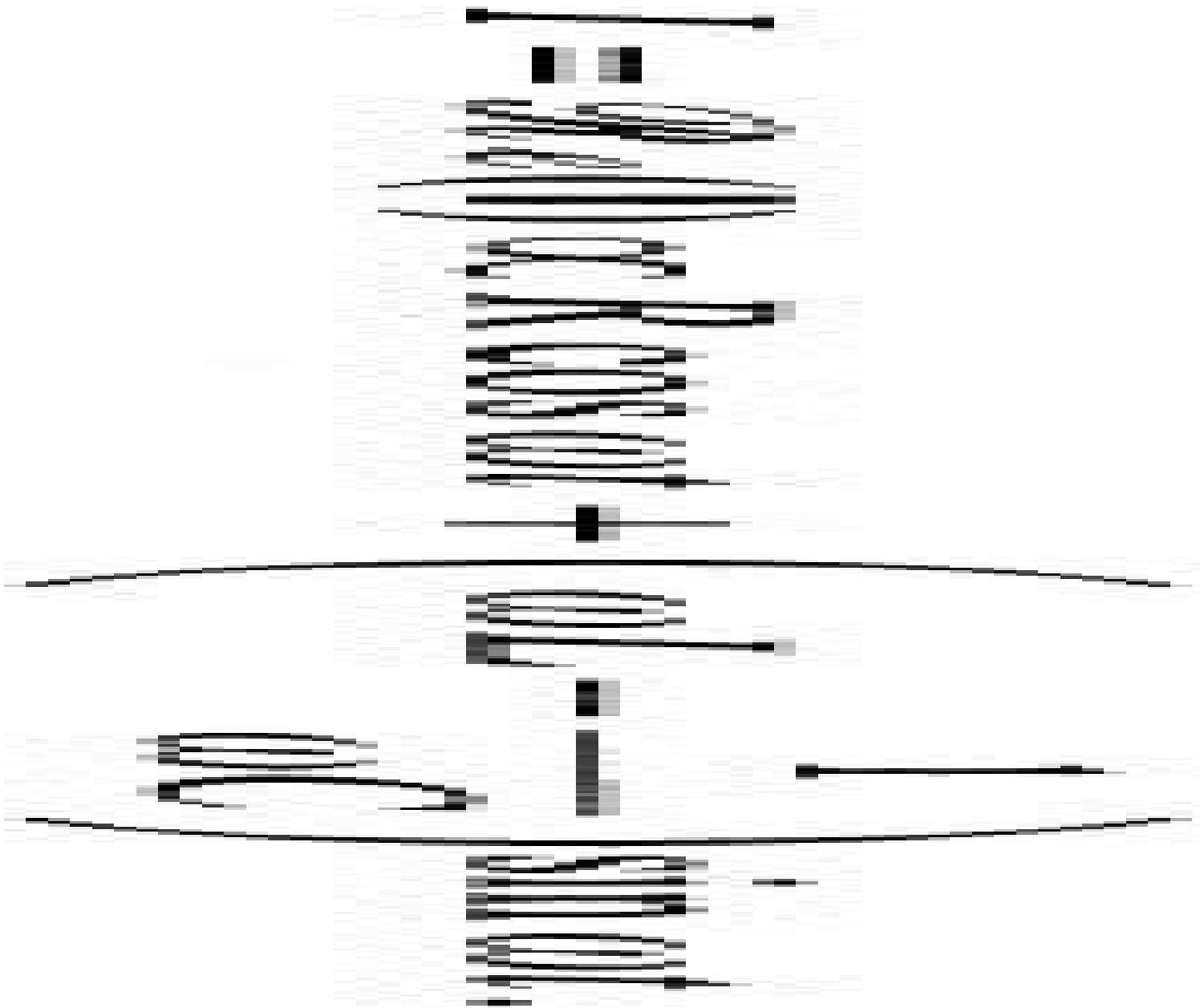
Discussion of complex impedance has allowed us to work out the behaviour of the current in the series LCR circuit. The same result can also be derived by direct solution of a differential equation for the current. Writing the voltage $V_0 \cos \omega t$ of the generator as the sum of the voltages across each of the components L, C and R, we have

(7.28)

where q is the charge on the capacitor at time t. Differentiating with respect to time, and replacing dq/dt by I,

(7.29)

The solution to this equation, as may be verified by substitution, is just the real part of the complex current already given in Equation (7.23):



Other resonance systems always satisfy a differential equation which has the same form as Equation (7.29) near resonance. For example, the rectangular cavities discussed in Chapter 12 act as very high frequency resonators, and the current they draw is related to the applied A.C. voltage by an equation like 7.29, even though the cavity is not made up of discrete components. The power absorption of a cavity peaks at resonance, and the amplitude and phase of the current vary in the same way as for the series LCR circuit. Energy losses occur because of currents flowing in the cavity walls and because of power returning along the coupling to the generator. The resonant frequency of a cavity is typically in the GHz region (1 GHz ≡ 10^9 Hz) and at such frequencies Q-values are very high. By using superconductors to minimize losses due to wall currents, Q-values of more than 10^8 can be achieved. Laser cavities operate at even higher frequencies. A Q-value of 10^9 is common for a laser, corresponding to a frequency spread of about 1 MHz at a visible light frequency near 10^{15} Hz.

A differential equation of the same form as Equation (7.28) describes resonance not only in electromagnetism but also in other branches of physics, and the solution always exhibits the same characteristics. Most familiarly, resonance occurs in mechanics. For example, consider a mass m attached to a spring with a force constant k, and damped by a velocity-dependent force γdx/dt. If the mass is now driven by a force F0 cos ωt, its position x is given by the solution of the differential equation

(7.30)

This equation has exactly the same form as Equation (7.28), and we see that the inductance L is analogous to the mass m, and the capacitance C to the reciprocal of the force constant k. Just as the inertia of the mass resists a sudden change of position, so the inductance prevents a sudden change in the amount of charge at any point in the circuit. The mechanical system has a resonance at its natural angular frequency The amplitude at resonance depends on the strength of the driving force and on the amount of damping. This is analogous to the resonance

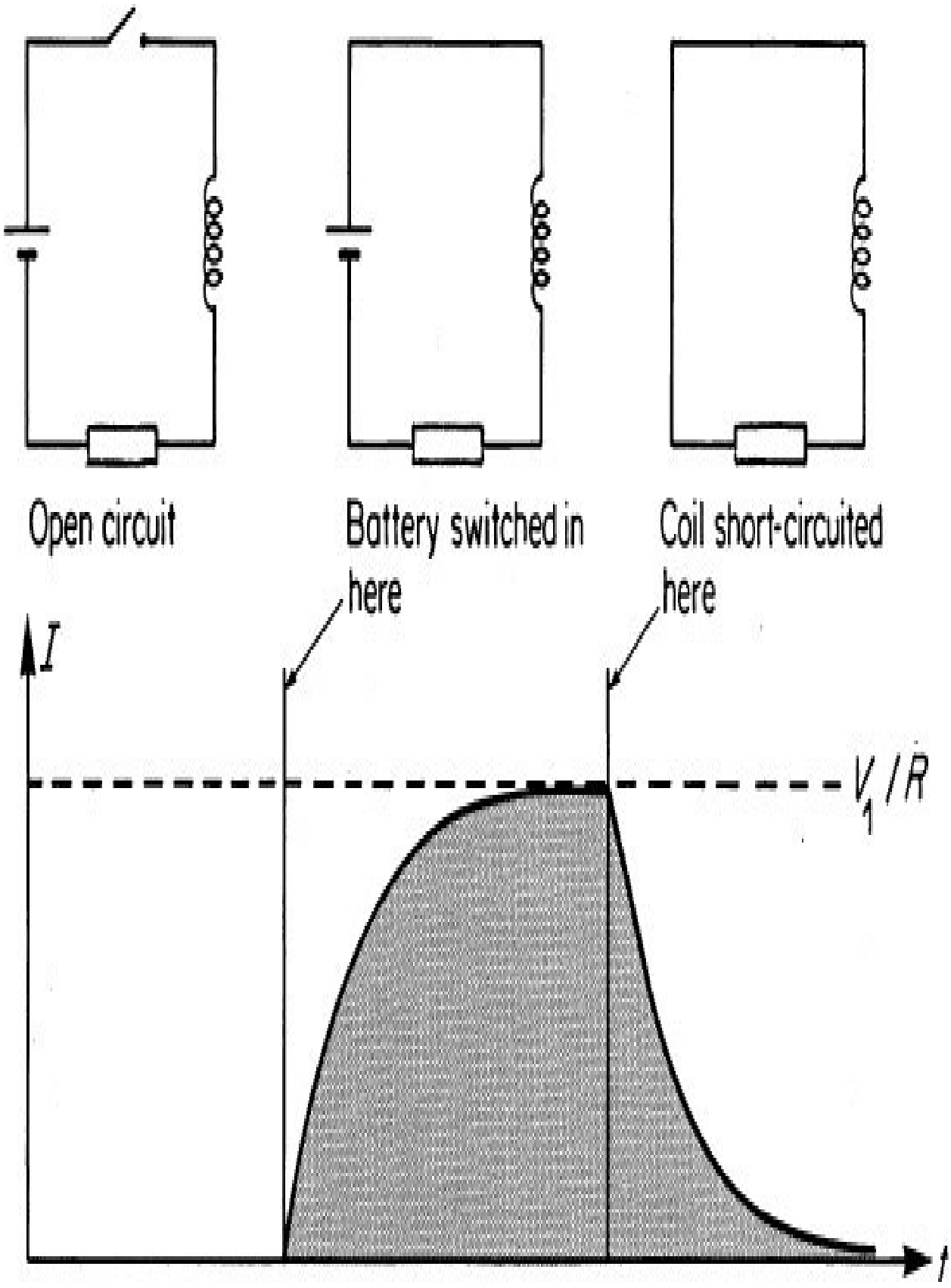
in the electrical circuit, which occurs at an angular frequency , with a current amplitude V0/R.

7.7 TRANSIENTS

Until now we have been considering the behaviour of electrical circuits in an equilibrium state, in which an A.C. generator maintains alternating currents of a fixed frequency everywhere in the circuit. Whenever a change is made to the e.m.f. in a circuit containing capacitance or inductance, transient currents flow after the change. Let us consider, for example, the circuit shown in Figure 7.11 consisting of a coil of self-inductance L and resistance R, and a battery of voltage V1. Initially the switch is open, and no current flows; the circuit is in an equilibrium state. When the switch is closed, current is driven through the circuit by the battery, and a new equilibrium state is approached in which the current is I = V1/R. But the current cannot suddenly jump to its new value, because a changing current through the coil induces an e.m.f. which opposes the change. Including the induced e.m.f., the differential equation relating current and voltage in the circuit is

(7.31)

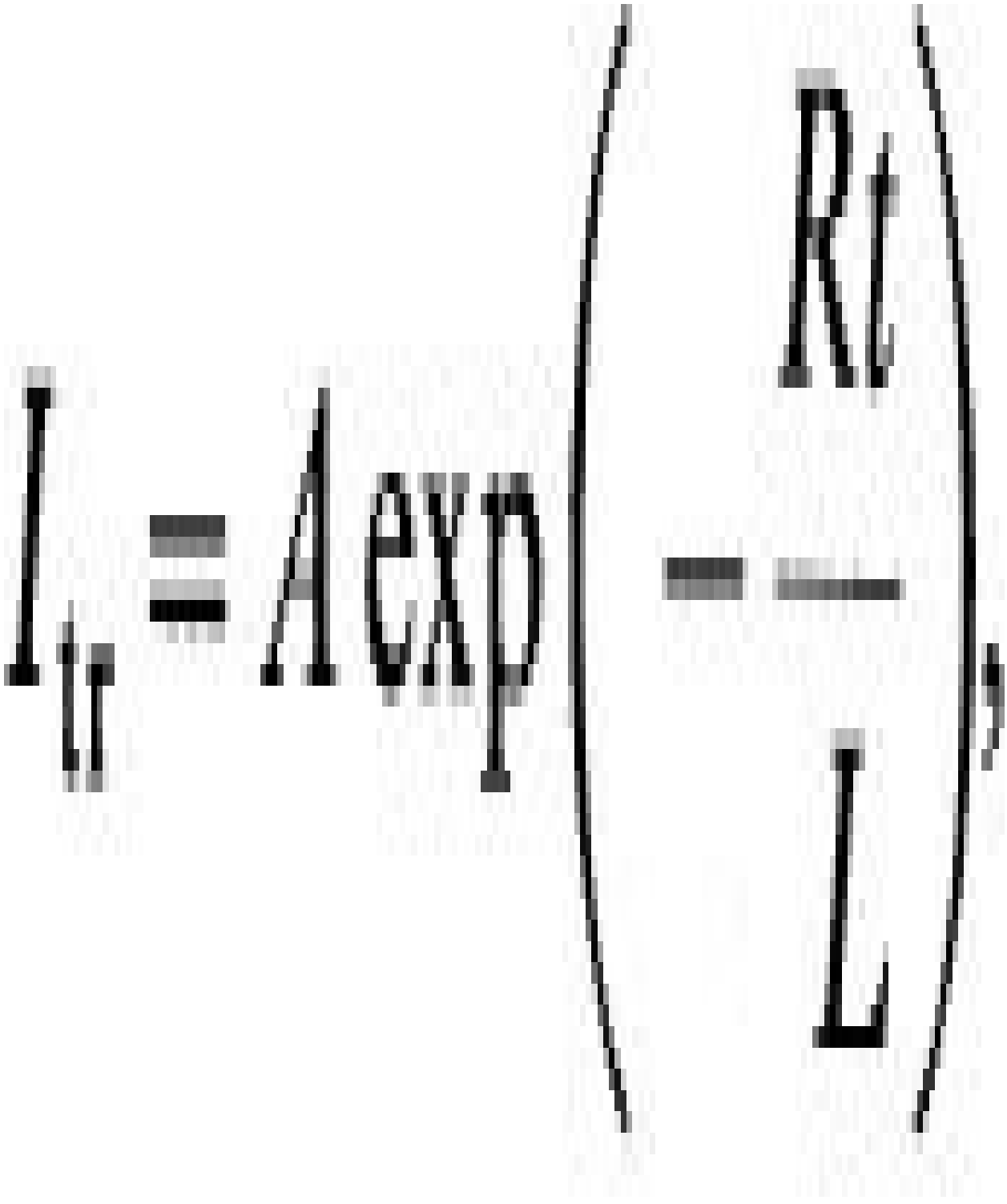
Figure 7.11. Transient currents flow in a coil whenever there is a sudden change in the voltage across its terminals.


Open circuit
Battery switched in here
Coil short-circuited here
I
V_1/R
t

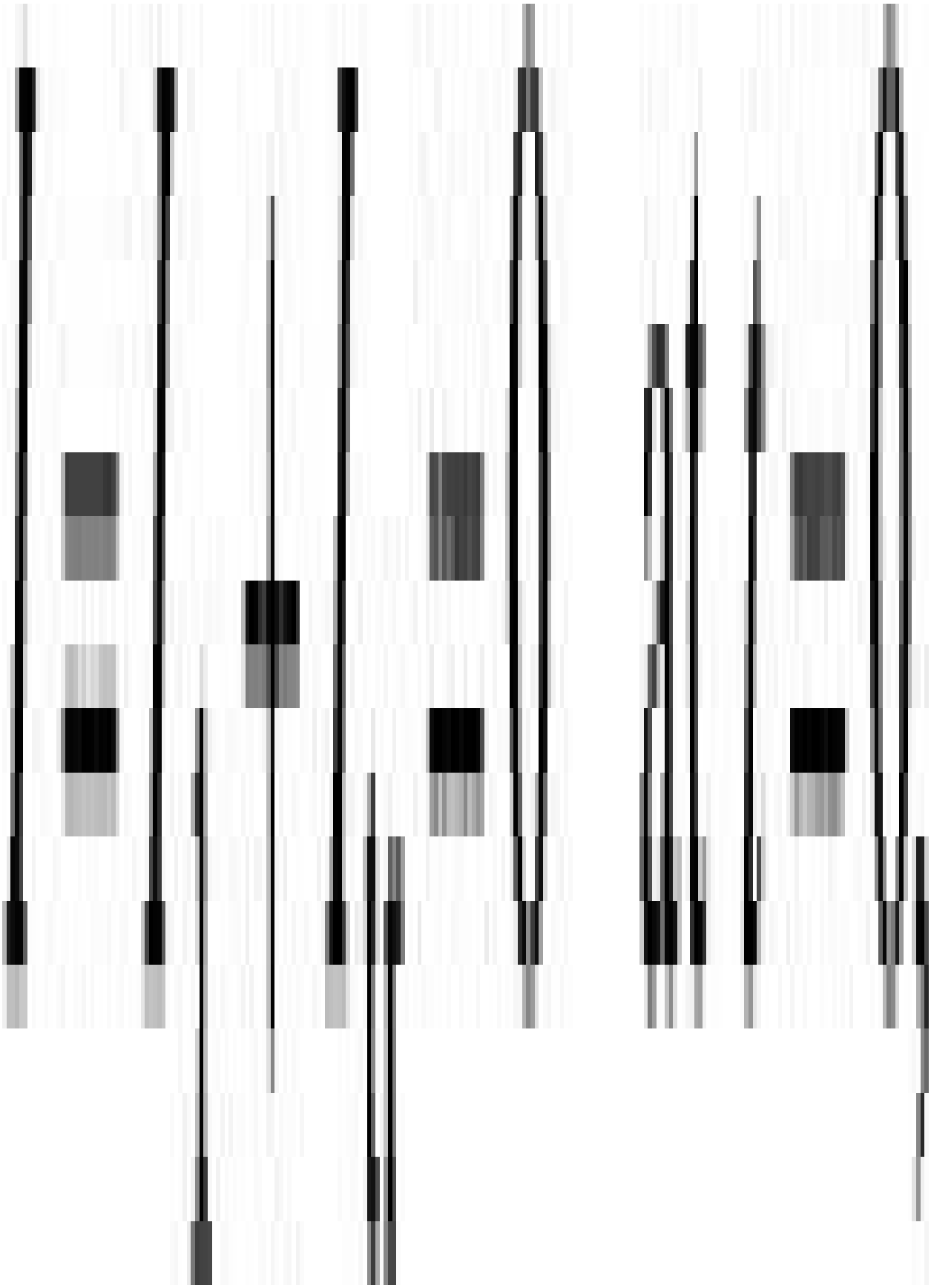
One solution of this equation is Ieq = I1 = V1/R, the constant current achieved at equilibrium. This solution is the particular integral of Equation (7.31). However, to the particular integral we can add any solution of the equation

(7.32)

The solution of Equation (7.32) is the complementary function of Equation (7.31); it is the complementary function which describes the transient currents in an electrical circuit. Writing the complementary function as Itr, we see by substitution in Equation (7.32) that

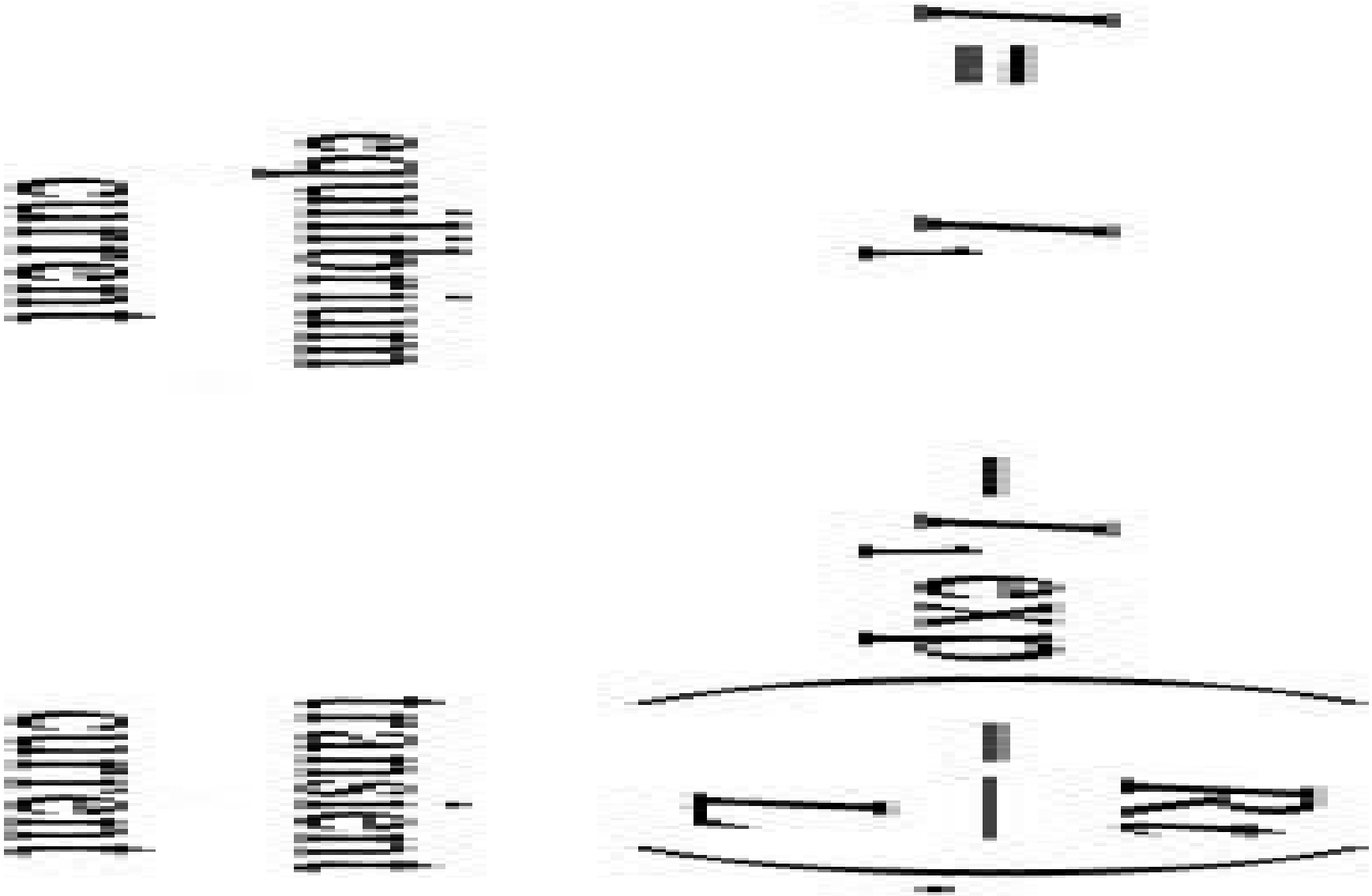


where A is any constant. The value of the constant is determined by the initial condition of the problem*, namely that the current is zero at the moment of switching. If we choose this moment to be the time-zero, then

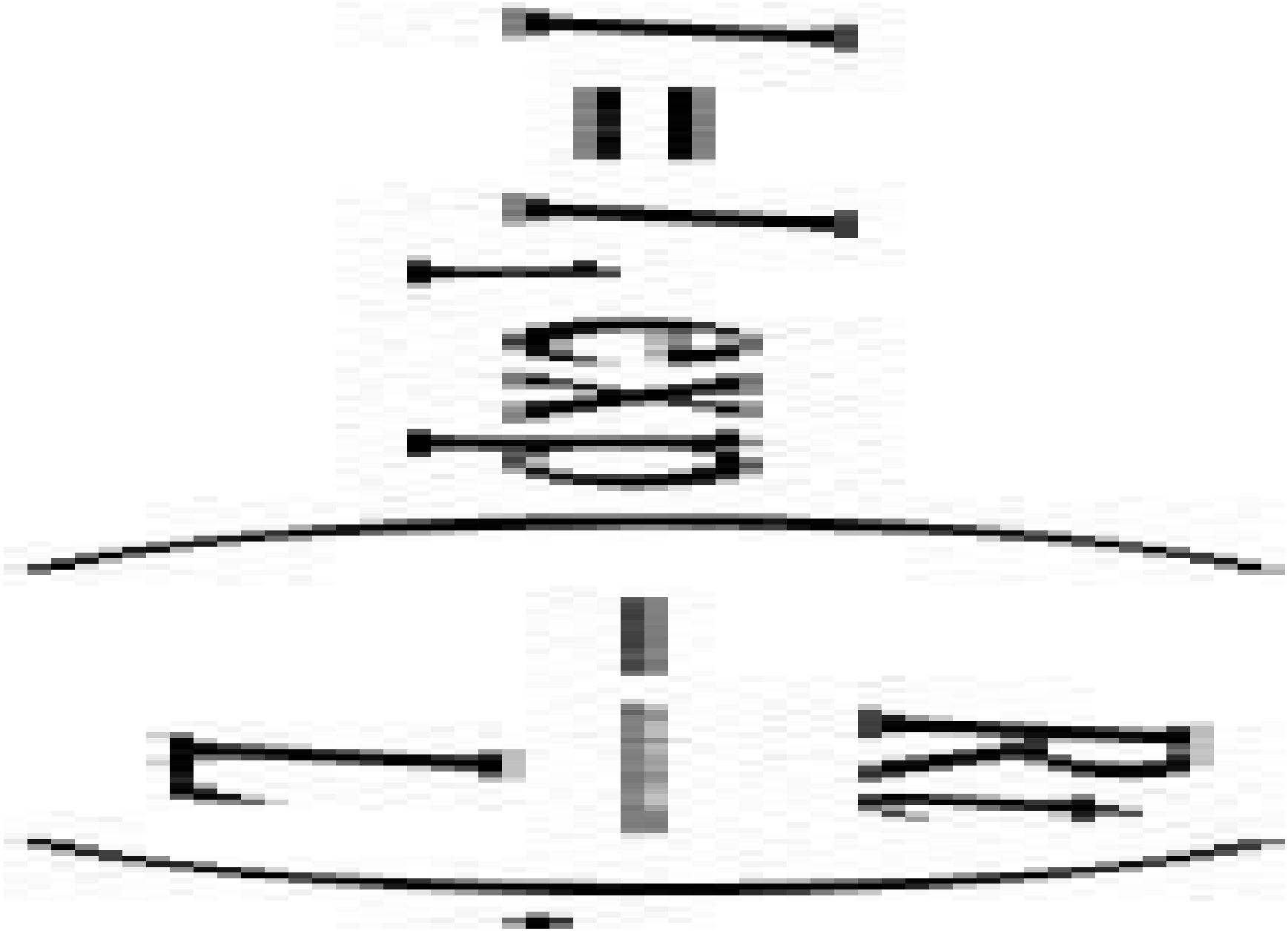


leading to A = –I1 and

(7.33)



In mathematical terms a transient current in any circuit is described by the complementary function of the corresponding differential equation. In terms of physical quantities, we can see that because reactive circuits store electrostatic and/or magnetic energy, there cannot be a sudden jump between two equilibrium states of the circuit. The transient current allows the changes to occur gradually, while energy is absorbed or delivered by the circuit. For example, as the current in a coil increases towards its equilibrium value according to Equation (7.33), the battery is doing work to build up the stored energy. The work done in a circuit containing self-inductance has already been discussed in section 6.3. There it was shown that energy is stored by the coil as the current builds up to its equilibrium value I1 Similarly, if the coil is suddenly short-circuited while carrying a current I1 as shown in Figure 7.11, a transient current flows until the stored energy is dissipated. The complementary function is the same as when the battery was switched into the circuit, and the transient current dies away exponentially with time constant L/R. Now the equilibrium current is zero, and since the initial current is I1 the current at a time t after short-circuiting is simply



As we showed in section 6.3, the rate of decay of the current in the shorted coil is such that heat losses in the resistance account for the whole of the energy initially stored in the coil.

Very rapid decay of the current in a coil can occur if it is suddenly open-circuited. Huge e.m.f.s are induced, causing a spark to cross the break in the circuit; most of the energy stored in the coil is dissipated in the spark. Because of the large e.m.f.s induced, it is dangerous to break the circuit of a big coil.

Sometimes transient currents do not die away monotonically, but show oscillations with a diminishing amplitude. Similar behaviour occurs in mechanical systems. Imagine that a mass on a lightly-damped spring is at rest. Now stretch the spring and then release it. The spring will oscillate at its natural frequency and gradually return to rest in its equilibrium position. If, however, the spring is heavily damped by being immersed in viscous oil, the mass returns slowly to its equilibrium without oscillating. In both cases the motion of the mass is described by the same differential equation, which is just Equation (7.30) with the force set equal to zero.

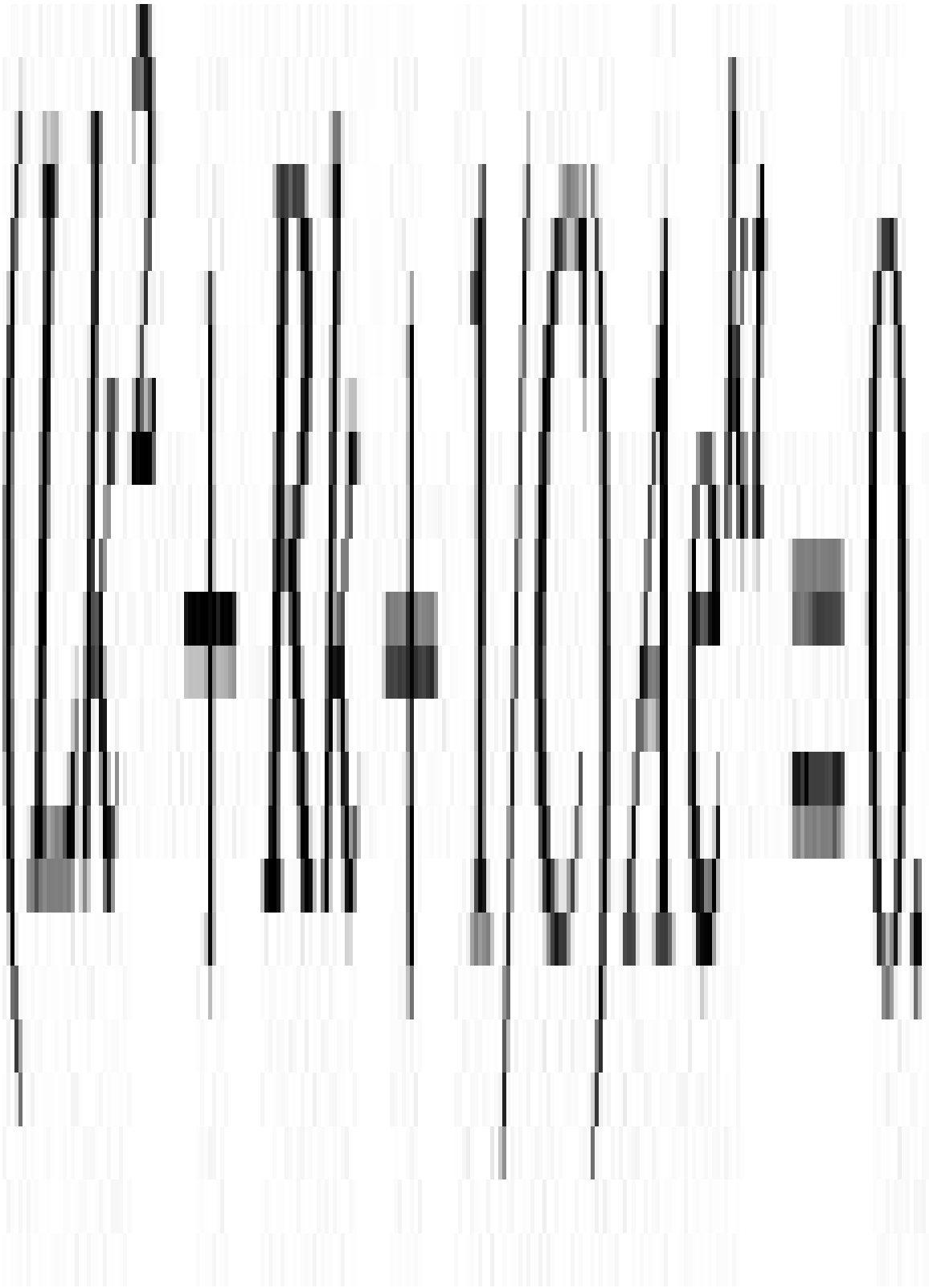
The electrical analogue is the series LCR circuit, with no A.C. generator. Such a circuit is shown in Figure 7.12. The circuit includes a battery, which determines the initial condition when the switch is closed. The differential equation for the charge q on the capacitor is the same as Equation (7.23) with the alternating voltage replaced by a steady voltage:

(7.34)

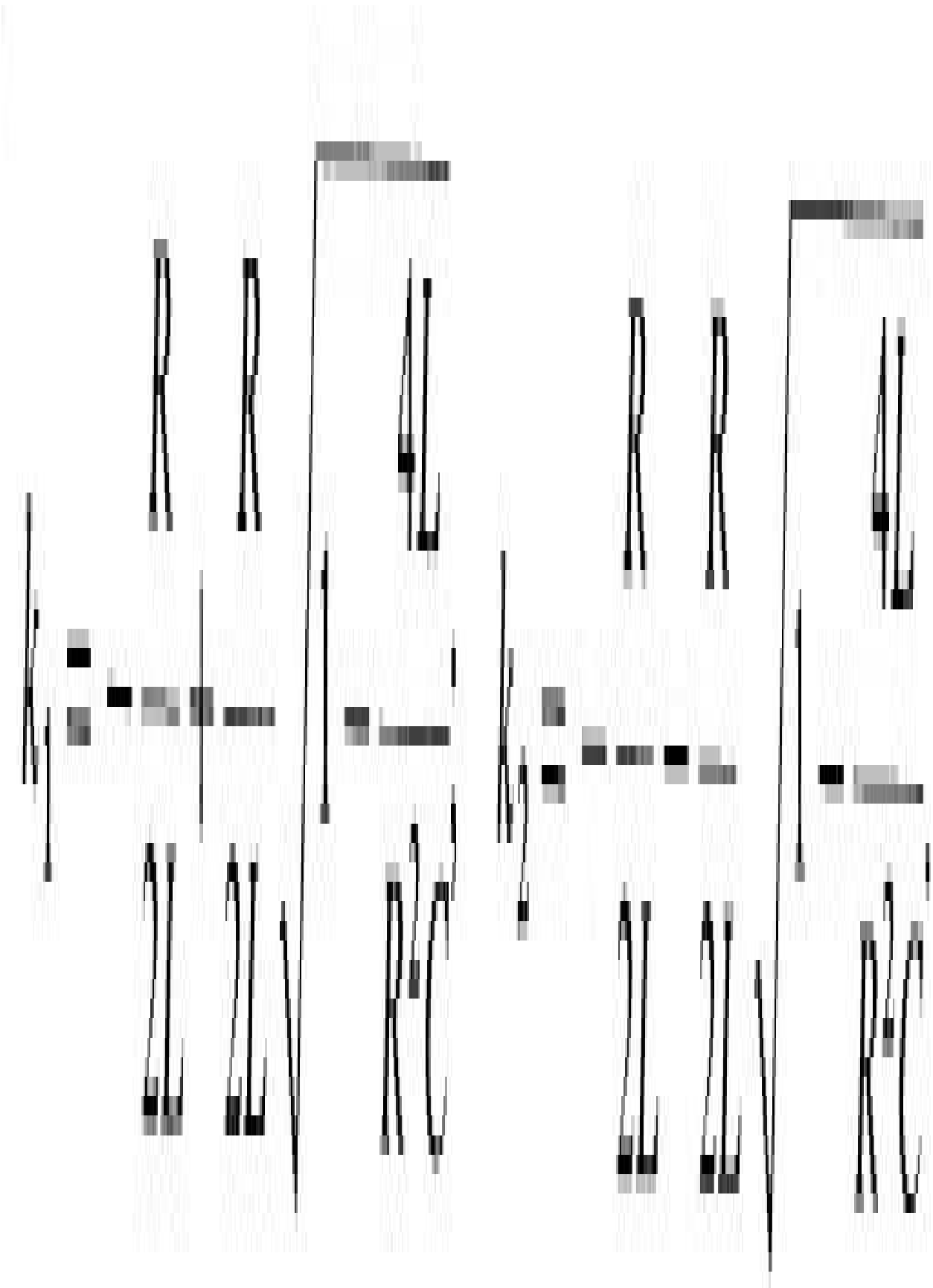
Differentiating with respect to time, and putting I = dq/dt,

(7.35)

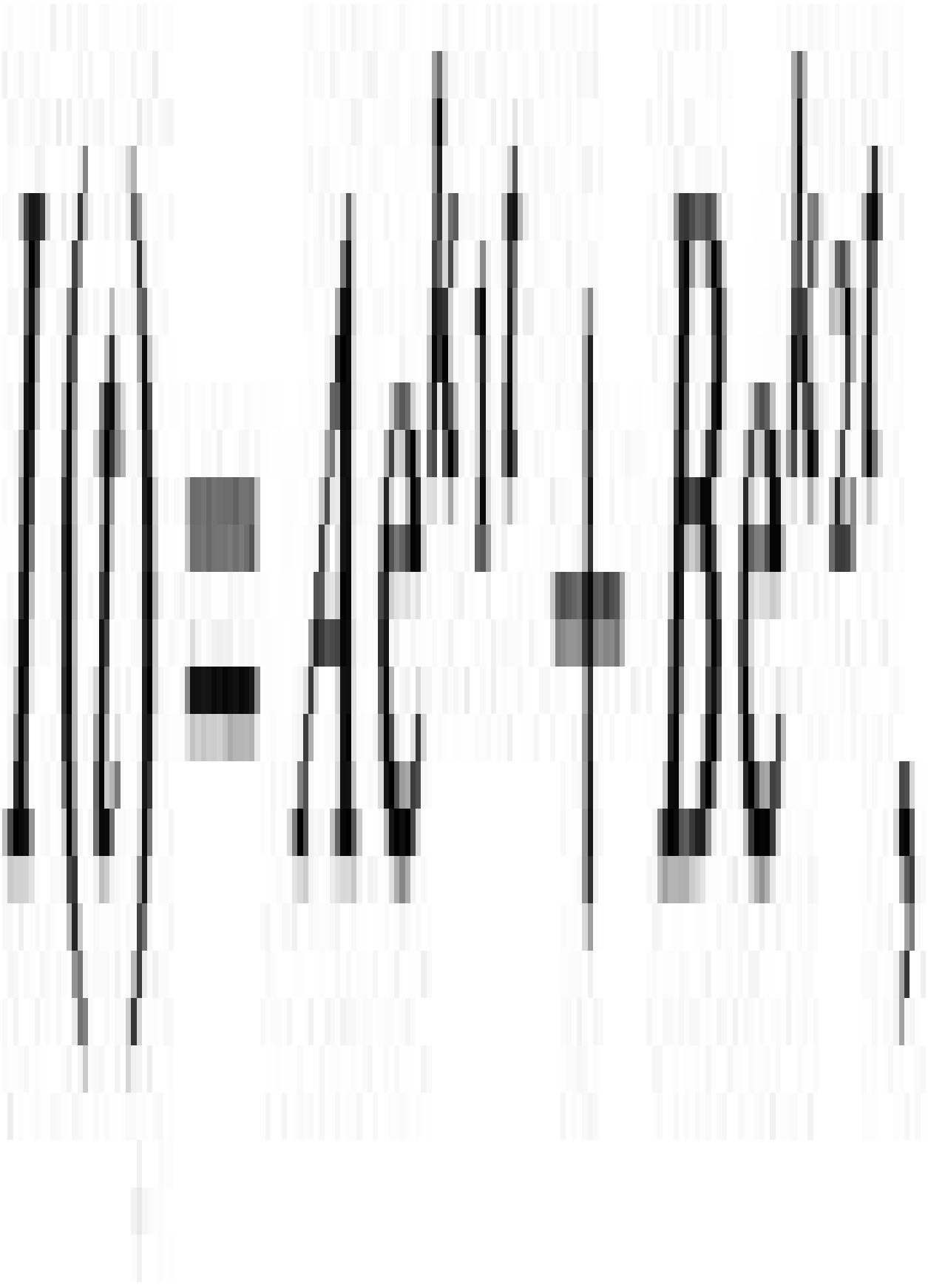
After the switch is closed, the circuit approaches an equilibrium in which the whole of the voltage appears across the capacitor and no current is flowing. By comparison with the mechanical analogue, we may anticipate that the transient may be oscillatory or not, depending on the value of the resistance. The differential equation may be solved by trying a solution of the form $I = Aekt$. This leads to



The exponent k must satisfy a quadratic equation, so there are two solutions



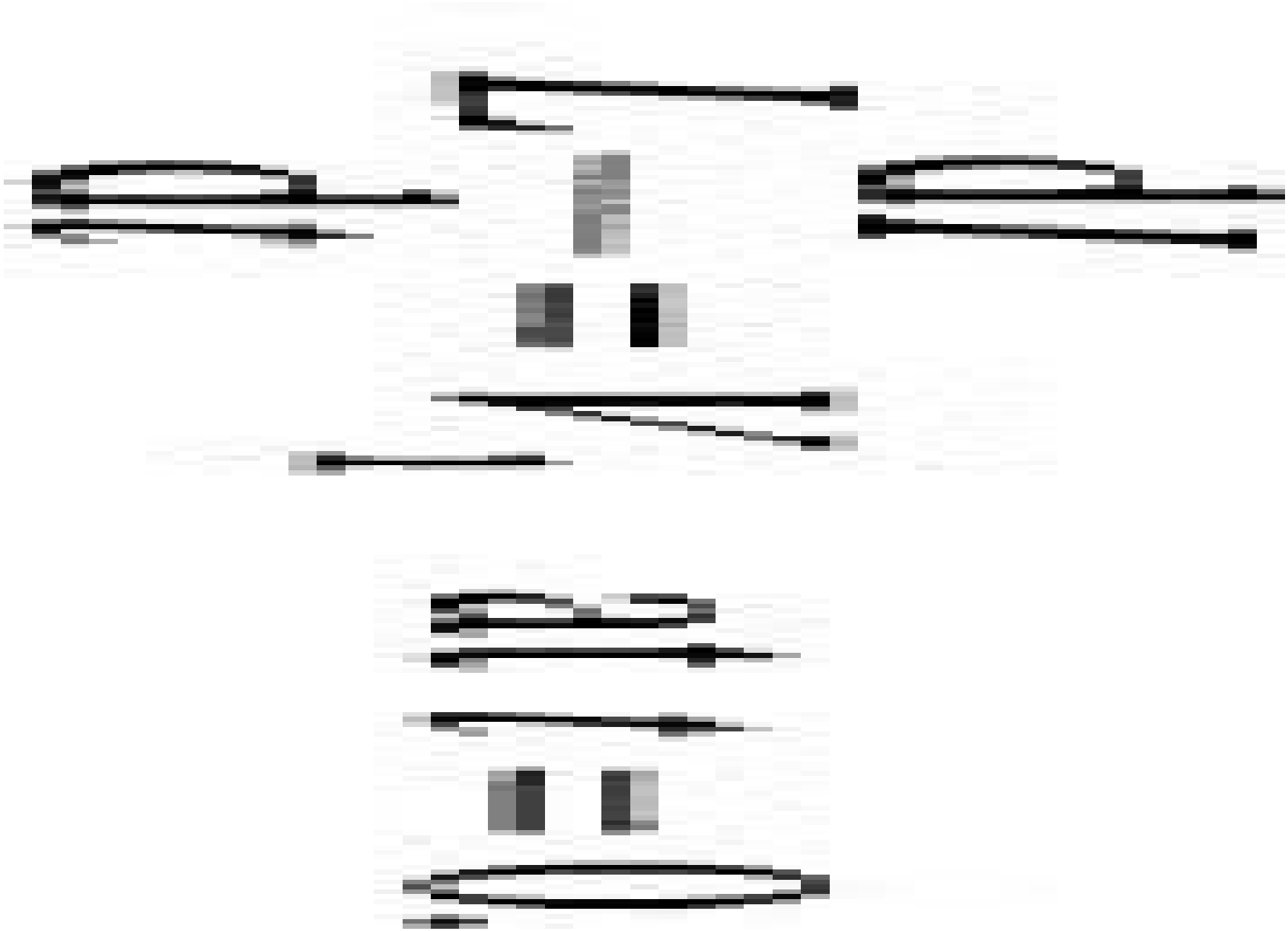
The general solution is then



where the constants A and B must satisfy the initial conditions The two conditions are:

(i) because of the presence of the inductance L, there can be no discontinuity in the current, and the initial current must be zero;

(ii) since the initial current is zero, at the moment of closing the switch the whole of the battery voltage appears across the inductance, hence



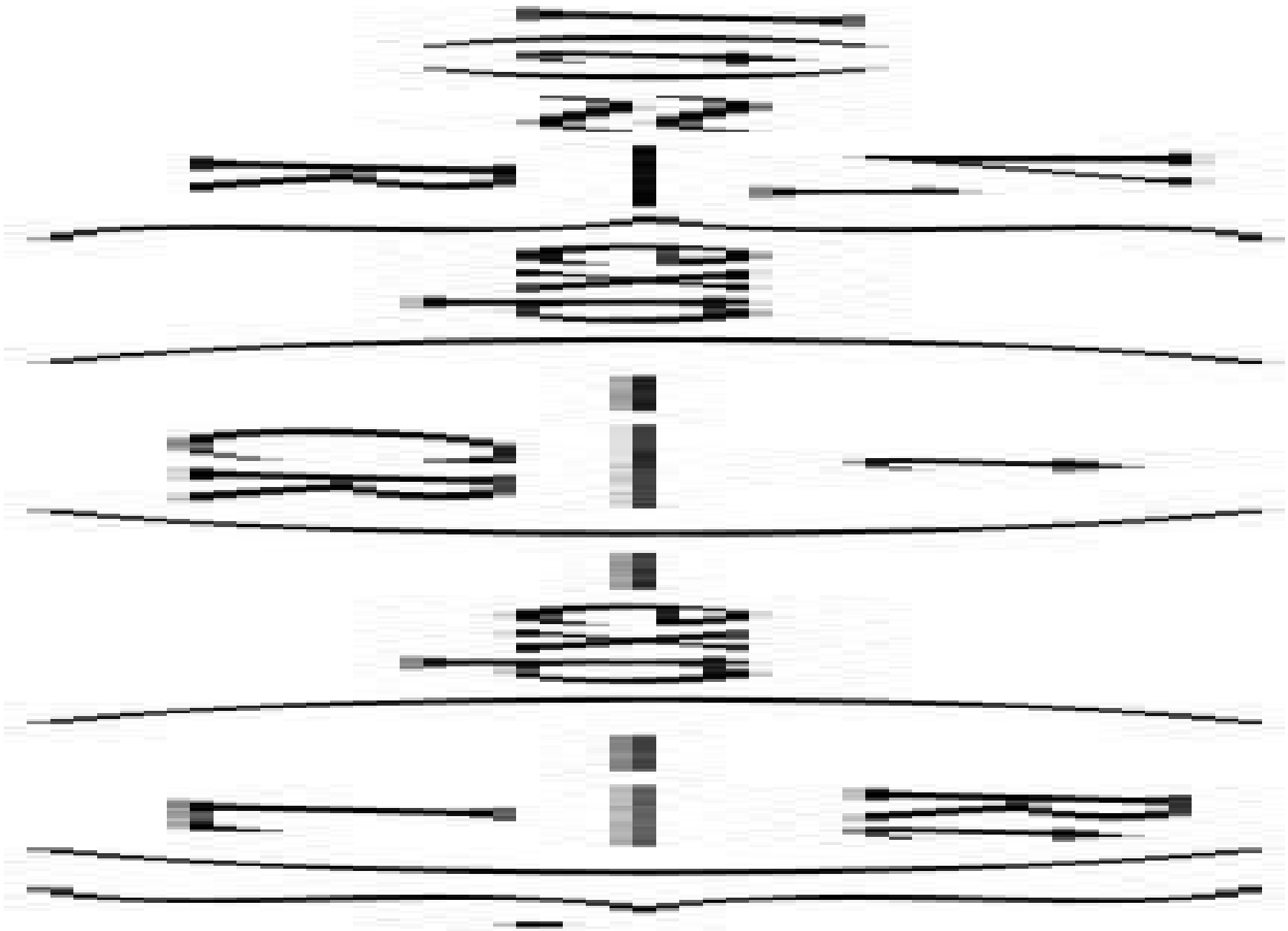
The current which satisfies these two initial conditions is

(7.36)

There are three different forms of this transient current, depending on the value of $4L/R^2C$.

(a) $4L/R^2C < 1$. Both k1 and k2 are then real and negative, so that the current is the difference of two decaying exponentials; the same time behaviour occurs in radioactivity when a daughter activity first builds up and then decays away again. This is the case of heavy damping, illustrated in Figure 7.12(a). If L/R^2C 1, Equation (7.36) has the approximate form

(7.37)



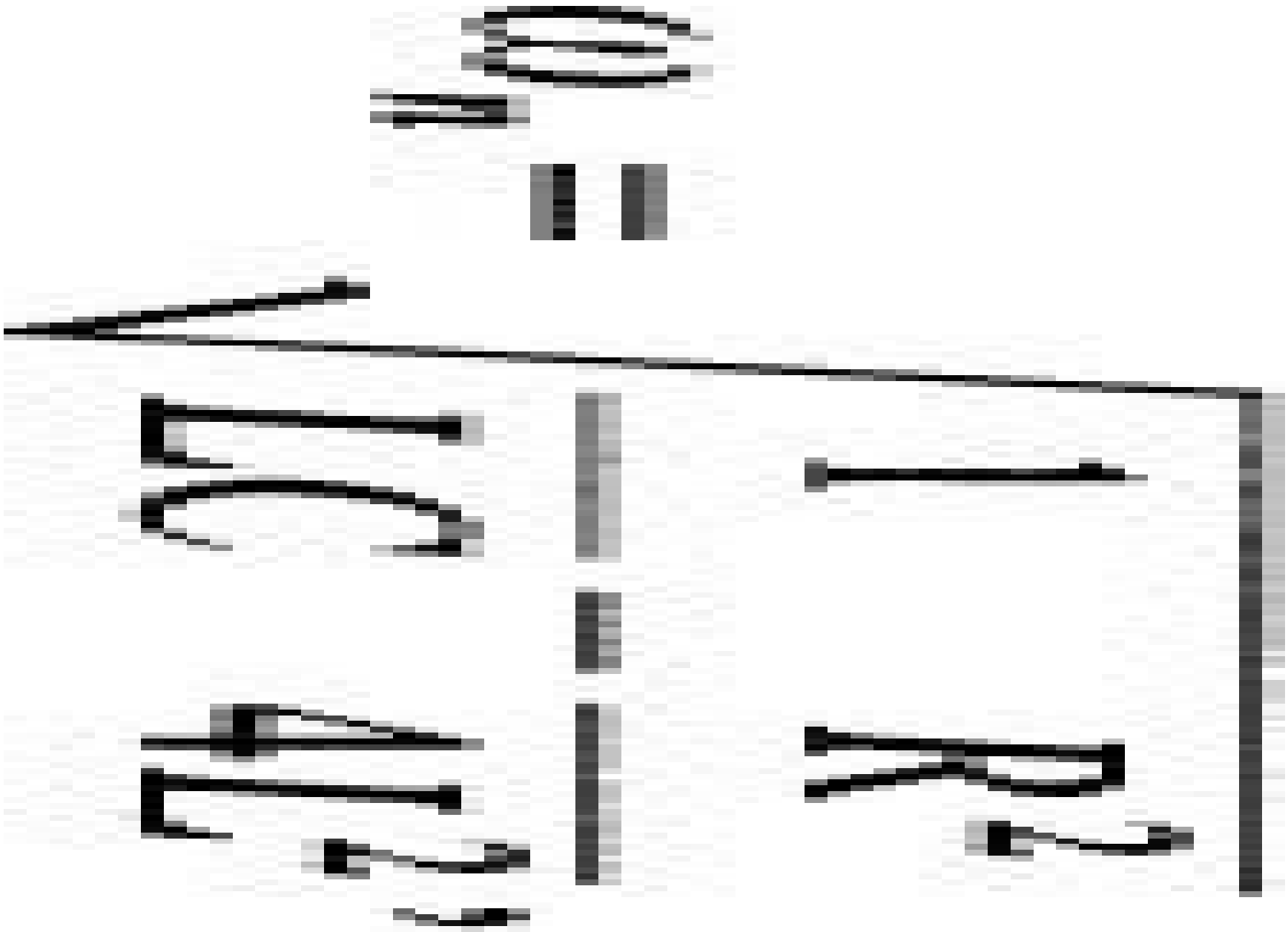
The time constant L/R RC, so the second term quickly dies away and the capacitor then charges up with the time constant RC.

(b) $4L/R^2C = 1$. In the limit as k1 approaches k2 Equation (7.36) reduces to

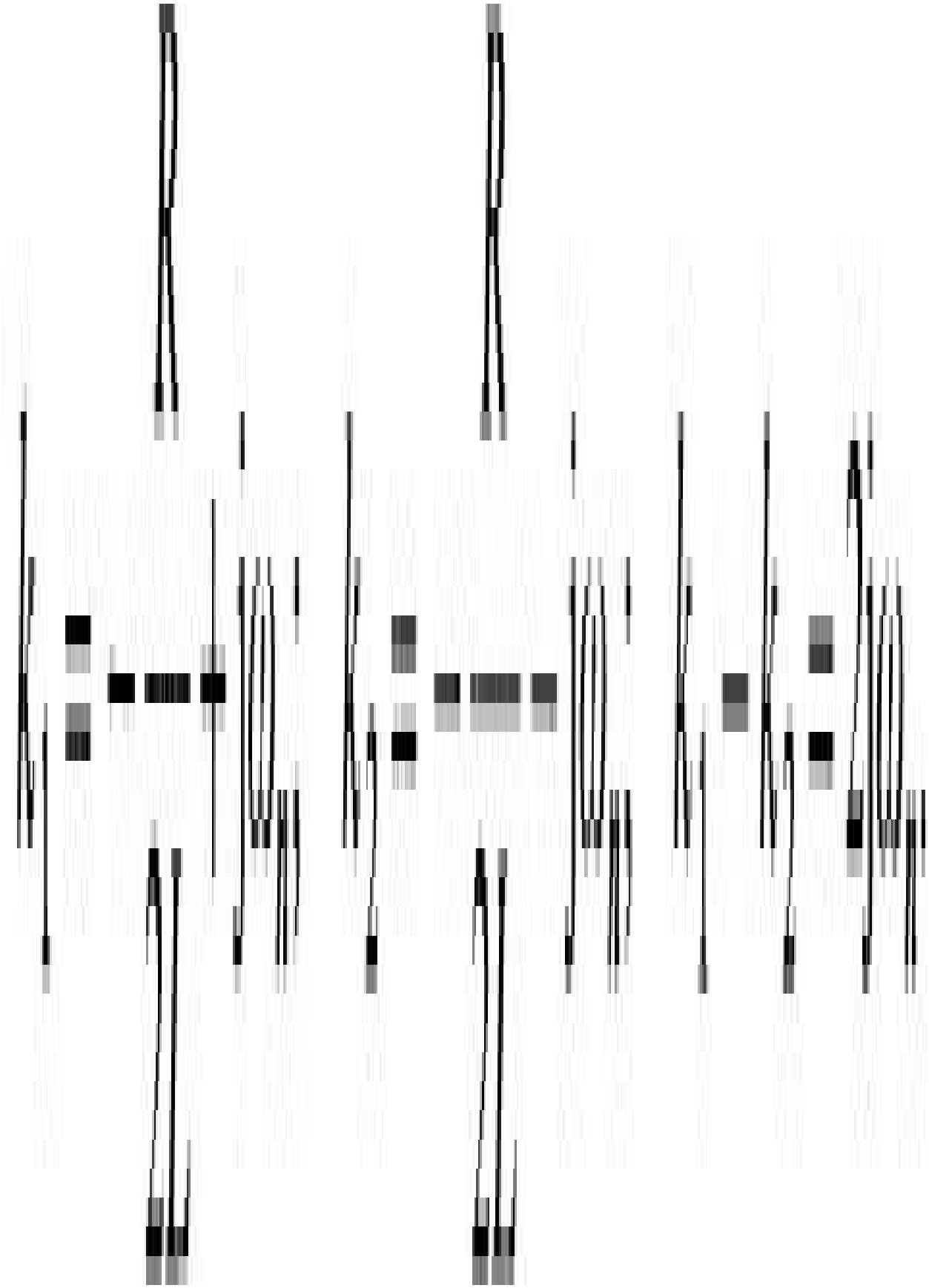
(7.38)

and the circuit is described as critically damped. For fixed values of L and C, the transient dies away most rapidly in a critically damped circuit. The critically damped transient current is sketched in Figure 7.12(b).

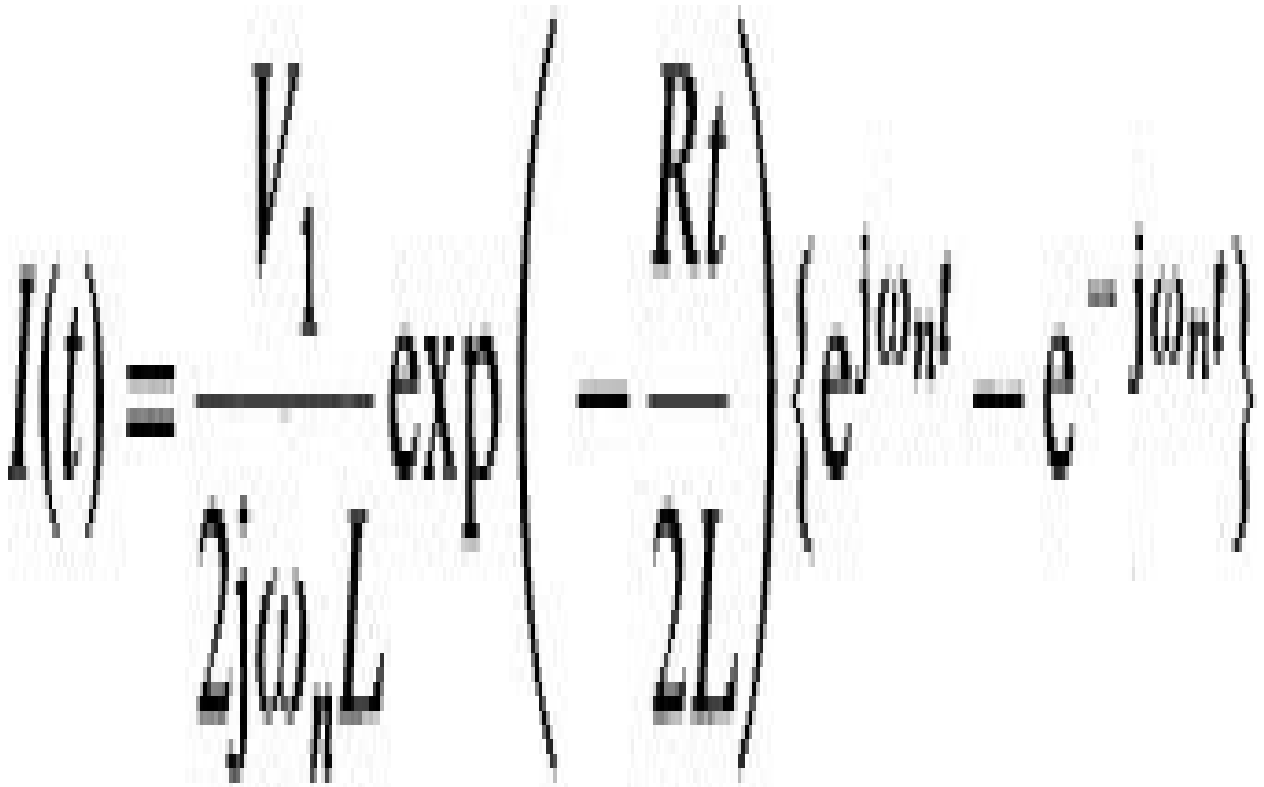
(c) $4L/R^2C > 1$. Under these conditions, k1 and k2 are no longer real. Let us write



so that



Equation (7.36) now becomes

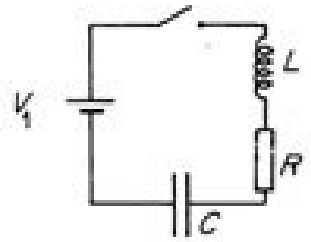


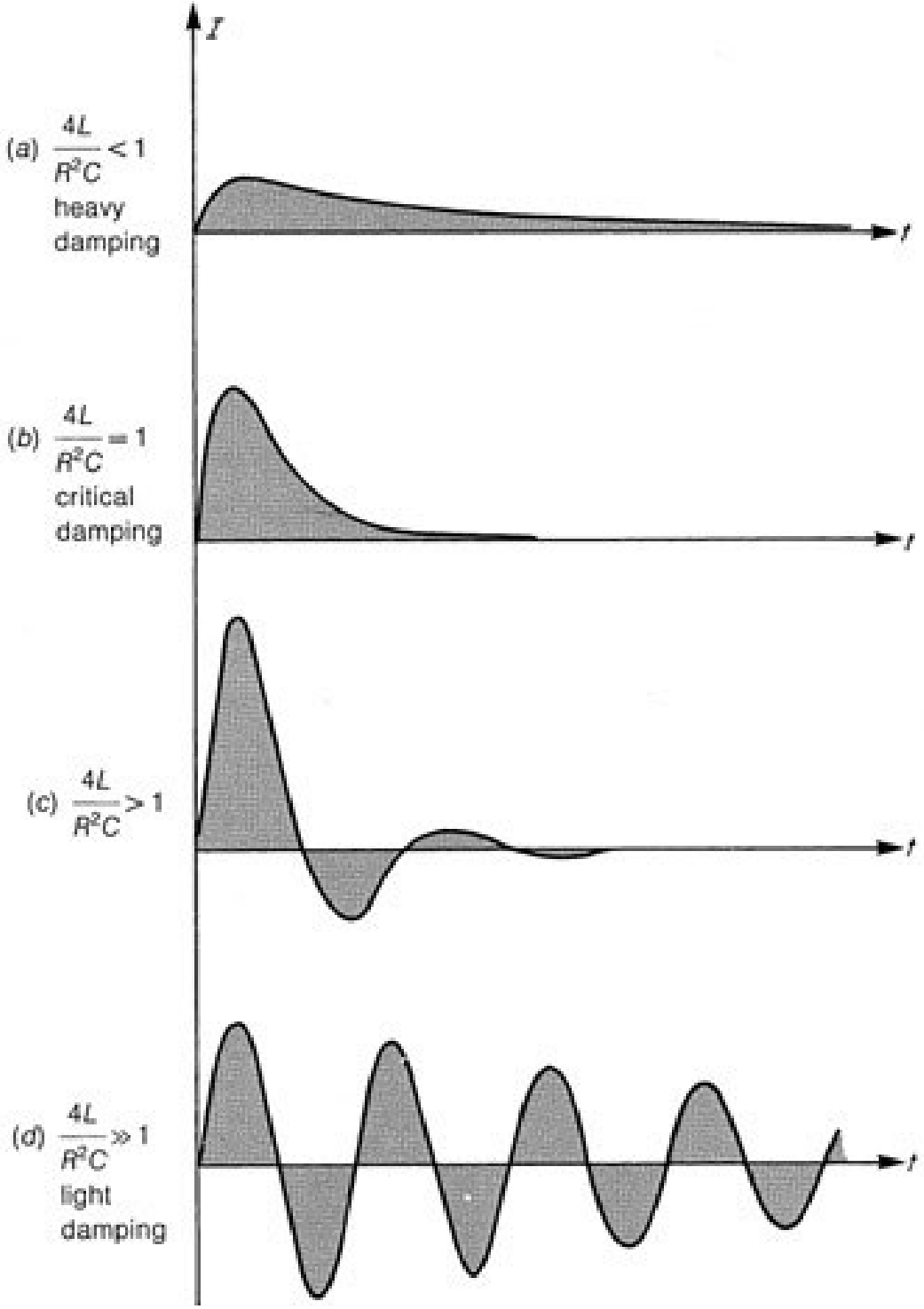
$$I(t) = \frac{V_1}{2j\omega_n L} \exp\left(-\frac{Rt}{2L}\right)\left(e^{j\omega_n t} - e^{-j\omega_n t}\right)$$

$$= \frac{V_1}{\omega_n L} \exp\left(-\frac{Rt}{2L}\right) \sin \omega_n t.$$

The current oscillates at the angular frequency ωn with an exponentially decreasing amplitude. Any circuit in which transients oscillate is said to ‘ring’, and ωn is called its natural angular frequency. Figure 7.12(c) illustrates ringing in a relatively heavily damped circuit in which the oscillations die out very quickly, and Figure 7.12(d) a lightly damped circuit for which $R^2/4L^2$ 1/LC. Under the latter condition the angular frequency of the oscillation is approximately equal to the resonant value . The amplitude of the oscillations gradually diminishes, and successive maxima, separated by the period 2π/ωn, are reduced by the factor exp(– πR/ωnL). The quantity πR/ωnL, which is called the logarithmic decrement of the circuit, is related to the sharpness of tuning, determined by Q = ω0L/R (see Equation (7.26)). We see that in a sharply tuned circuit, for which Q is large, the logarithmic decrement is small, and consequently transient oscillations take a long time to die away.

Figure 7.12. Transient currents in an LCR circuit as R changes for fixed L and C.


L
V_1
R
C


I
(a) $\frac{4L}{R^2C} < 1$
heavy
damping
t
(b) $\frac{4L}{R^2C} = 1$
critical
damping
t
(c) $\frac{4L}{R^2C} > 1$
t
(d) $\frac{4L}{R^2C} \gg 1$
light
damping
t

In spite of the different shapes of the plots in Figure 7.12, the final result is in each case the same. The battery has delivered a total charge CV1 to the capacitor, i.e. the algebraic area under all the current curves is equal to CV1, and the battery has delivered an amount of energy electrostatic energy is stored on the capacitor, and an equal amount of energy has also been dissipated in the resistance. Further, if the battery were replaced by a short-circuit, the stored energy would be dissipated as heat during a transient similar in form to the one occurring when the battery is switched into the circuit. Once again the transient has been shown to adjust the circuit to the requirements of stored energy in the reactive components.

PROBLEMS 7

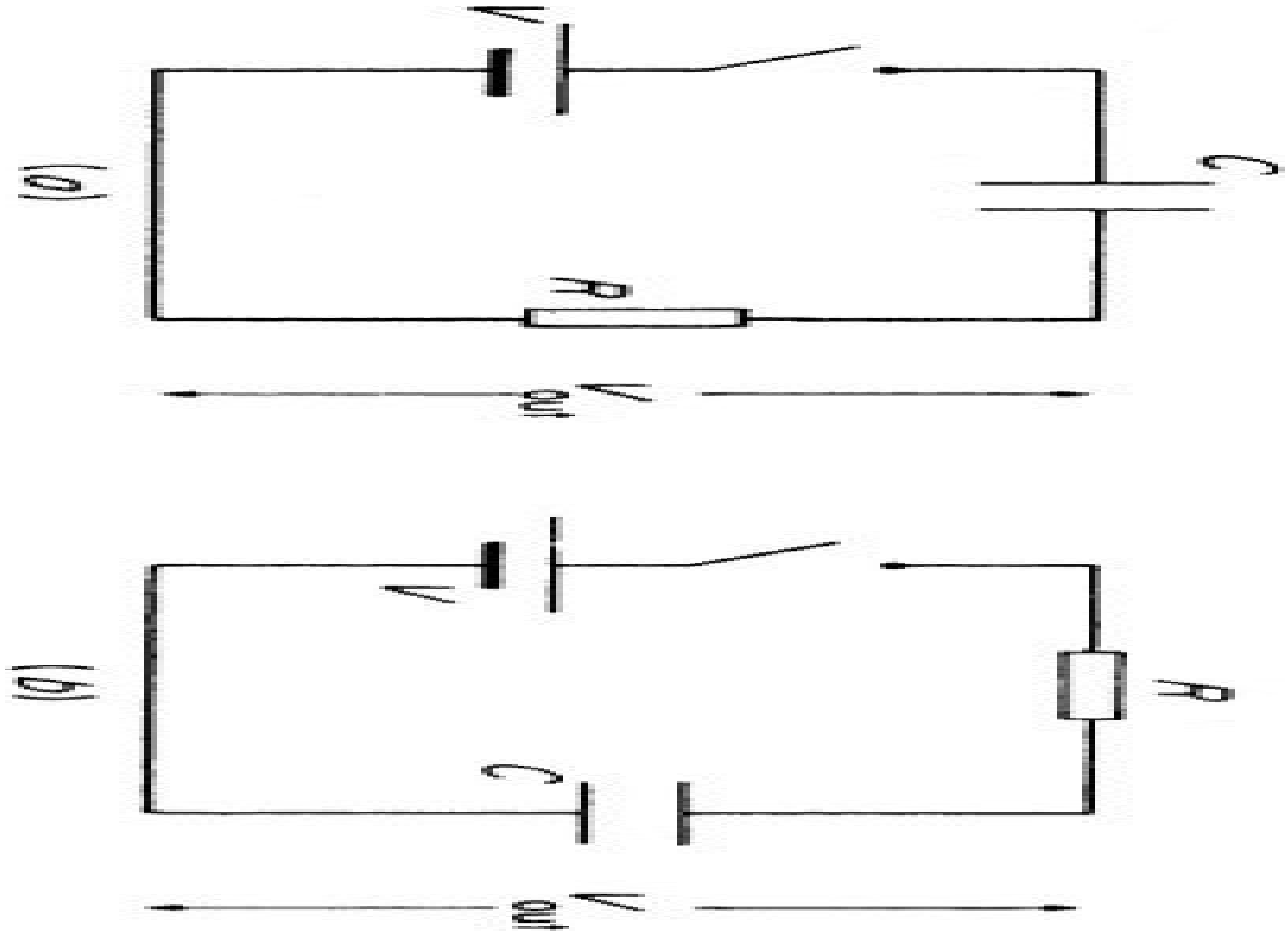
7.1 Calculate the amplitude and phase of Vout in the circuits shown in Figure 7.13.

Figure 7.13.

C
$V_0 e^{j\omega t}$
R
V_{out}
(a)
R
$V_0 e^{j\omega t}$
C
V_{out}
(b)

7.2 Sketch the voltage Vout as a function of time after the switch is closed in the circuits shown in Figure 7.14.

Figure 7.14.



7.3 A capacitor discharges through its leakage resistance with a time constant of 1 sec. What is the power factor of this capacitor when it is in a 50 Hz A.C. circuit? (You may represent it by an ideal capacitor in parallel with a resistor).

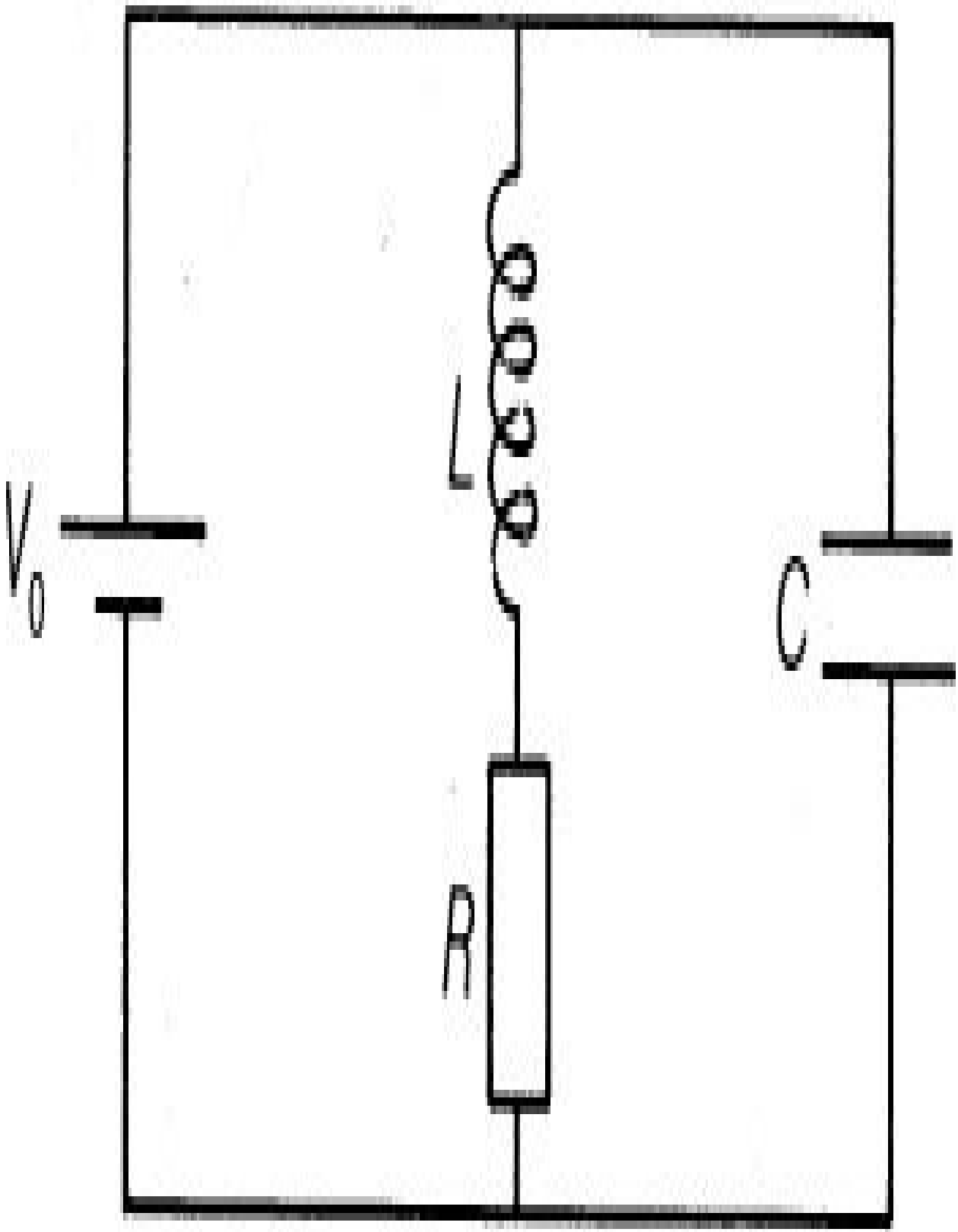
7.4 A 1 μF capacitor is placed in series with a 1000 Ω resistor. At what frequency is their impedance 2000 Ω? Draw the phasor diagram at this frequency.

7.5 The resistor and capacitor of the previous question are placed in series across a 240 volt 50 Hz mains supply. What is the RMS current drawn from the mains? What is the RMS voltage across the capacitor?

7.6 A coil has a self-inductance of 10 mH and a resistance of 100 Ω. What capacitance must be put in series with the coil to make a critically damped circuit? If the critically damped circuit is driven by an alternating voltage of constant amplitude, at what angular frequency is the current through the circuit a maximum? What is the Q of the circuit?

7.7 After the circuit in Figure 7.15 has reached a steady state the battery is disconnected, and the voltage across the capacitor is found to ‘ring’. If L = 1 mH, C = 0.1 μF and R = 10 Ω what is the natural frequency of the circuit, and what is its logarithmic decrement? Sketch the variation of the voltage across the capacitor after the battery has been disconnected.

Figure 7.15.


V_0
L
C
R

*The skin effect is discussed in Chapter 11.

*The problem of working out the value of the complex impedance of a network is discussed in Chapter 8.

*A full treatment of resonances in different systems can be found in Vibrations and Waves by A. P. French (Van Nostrand Reinhold, 1982).

*There is an alternative method of solving transient problems which avoids the cumbersome procedure of finding the particular integral and the complementary function separately, and then combining them to satisfy the initial conditions. This method, which is called the Laplace transform method, incorporates the initial conditions from the start. A good account of it may be found in Partial Differential Equations for Scientists and Engineers by G. Stevenson, 3rd edition (Longman, 1985).

CHAPTER 8

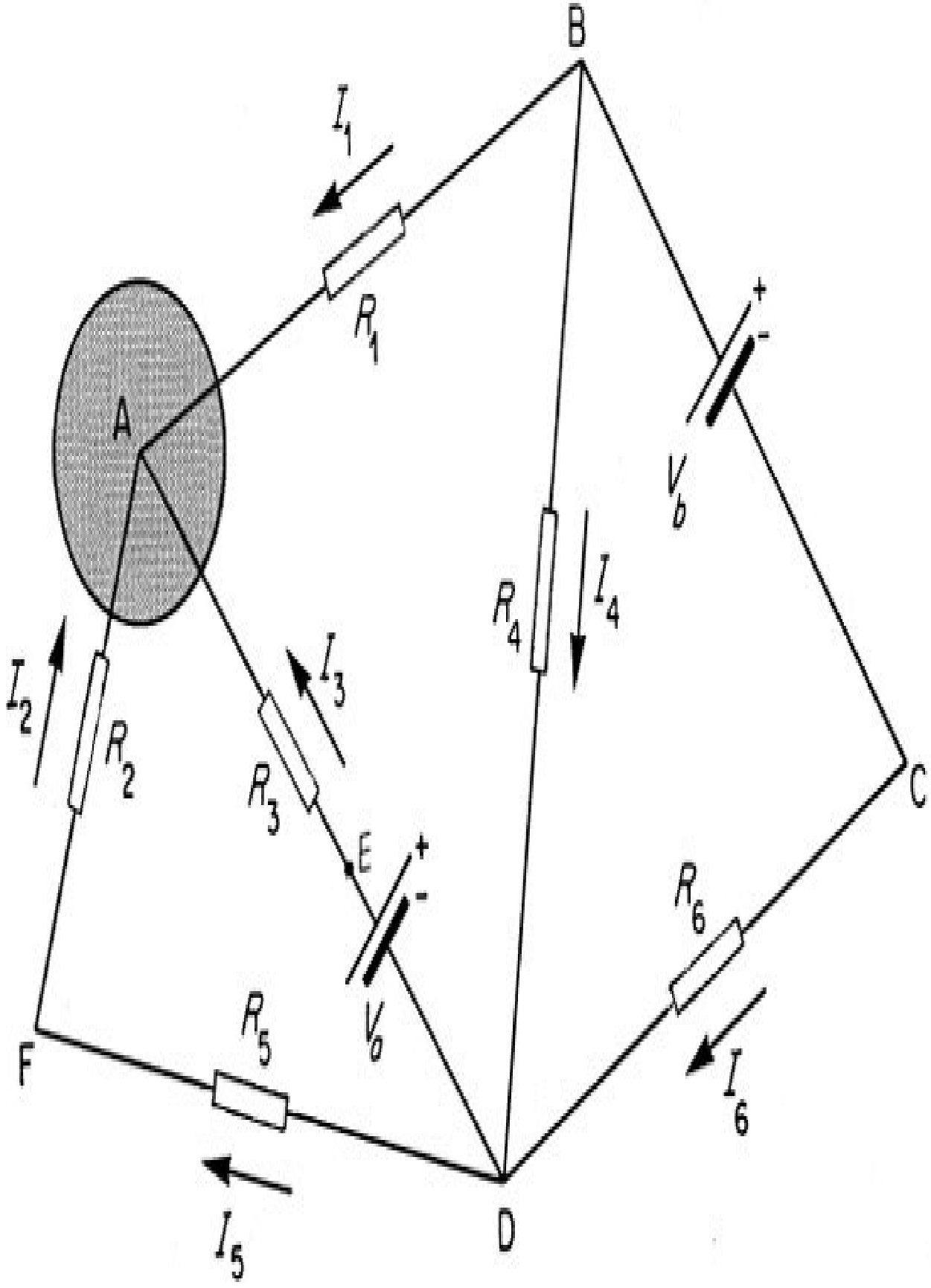
Linear circuits

A practical electrical circuit often contains many components. In order to predict the behaviour of the whole circuit, one must be able to calculate the current flowing through each component. The technique for doing this is known as network analysis. The first step in the analysis is to represent the real circuit by an equivalent circuit made up of idealized circuit elements which have almost the same behaviour as the components in the real circuit. We have already met this idea in discussing capacitors and coils in section 7.4. A coil, for example, is represented by an idealized inductance in series with a resistance. Although resistance is actually distributed along the length of the coil, the equivalent circuit behaves in the same way as the real coil. Similarly, the properties of a more complex circuit must be mocked up as closely as possible by an equivalent circuit containing only ideal elements. Such an equivalent circuit is known as a network. Once the network has been set up, two rules, known as Kirchhoff's rules, are sufficient to specify the currents flowing in all the elements, however complicated the network may be. In this chapter Kirchhoff's rules are explained, and they are applied to linear circuits. By linear circuits we mean those in which the current carried by each element is proportional to the voltage difference across it, either for direct current (D.C.) or alternating current (A.C.) at a fixed frequency.

8.1 NETWORKS

nd the resistors A network consists of any number of components linked together by connecting wires, which are considered to have zero impedance. Figure 8.1 is a diagram representing a D.C. network, made up of two batteries generating e.m.f.s Va and Vb, and the resistors R1, R2,…, R6. We shall use this figure to illustrate the meaning of a number of terms which are used in the discussion of networks.

Figure 8.1. A direct current network.


B
I_1
R_1
A
V_b
R_4
I_4
I_2
R_2
I_3
R_3
C
E
R_6
V_b
R_5
F
I_6
I_5
D

A node is a junction where two or more connecting wires are joined together: the points A, B, C, D, E and F in the figure are nodes.

A branch is the path through a component joining two nodes.

A loop is a closed path following the branches of the network: the paths ABDFA and BCDB are thus both loops.

8.1.1 Kirchhoff's rules

The currents in the components of a network, and the voltages across them, must satisfy Kirchhoff's rules, which are:

1. Kirchhoff's current rule

The net current arriving at any node of the network is zero;

2. Kirchhoff's voltage rule

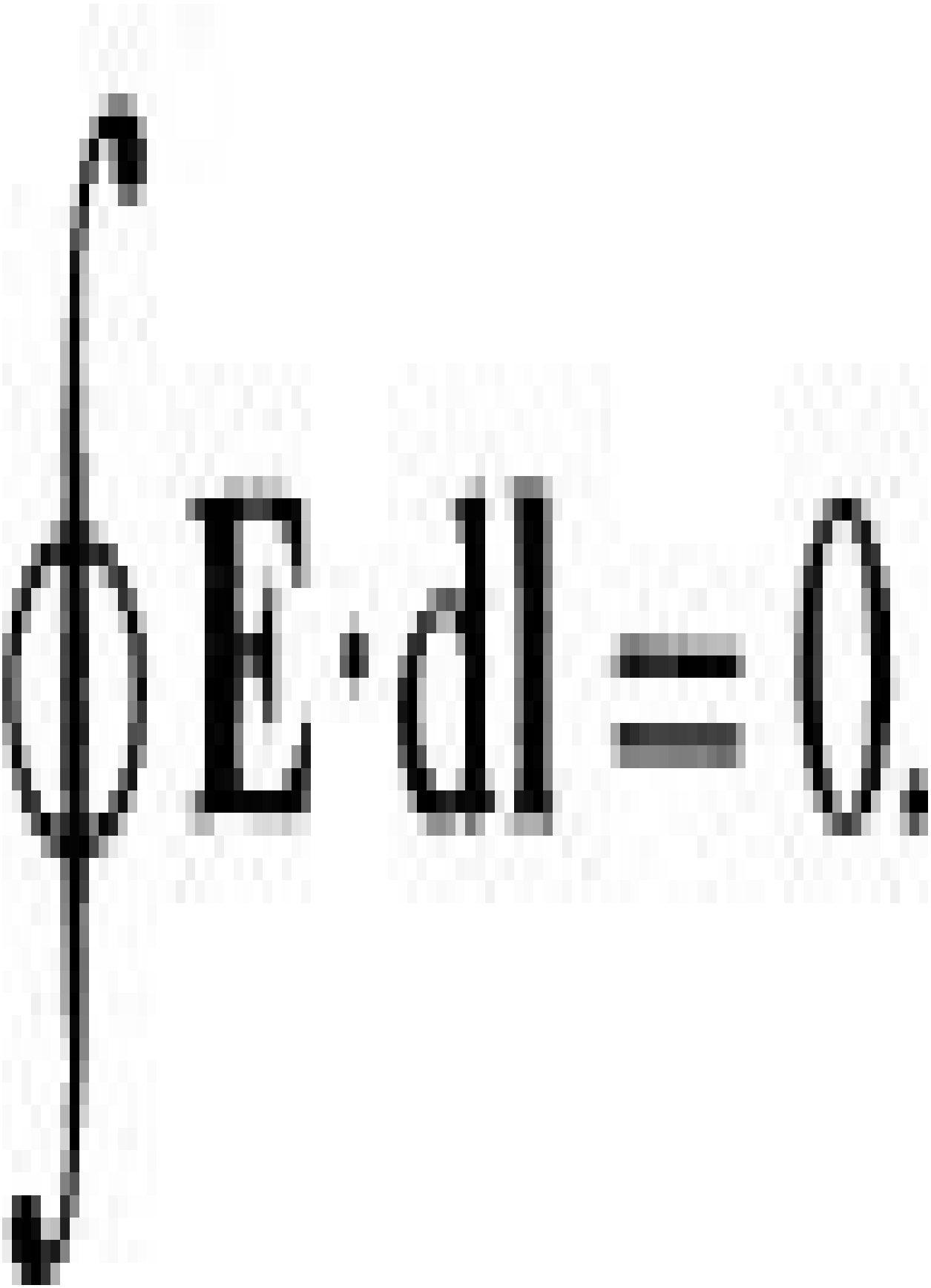
The sum of the voltage around a loop of a network is zero.

The rules are valid for both D.C. and A.C. circuits, but we shall justify them first for direct currents and later generalize to alternating currents.

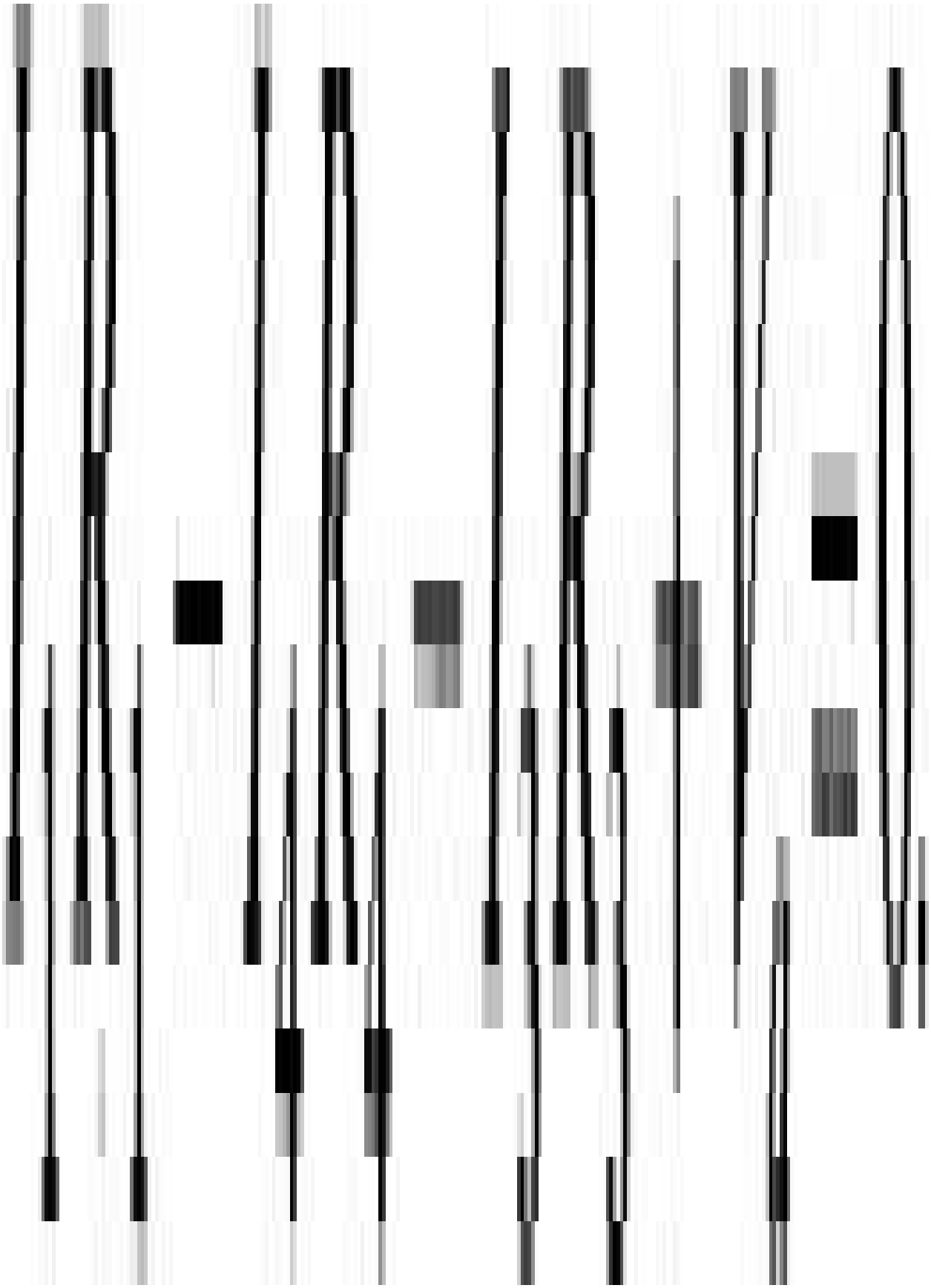
The current rule is an expression of the conservation of charge. In the figure, for

instance, no charge can accumulate at the node A when the network is in a steady state, and the total current entering the shaded area is $I1 + I2 + I3 = 0$. To satisfy this condition at least one of the currents must be negative. A negative value indicates that the currrent flows in the opposite direction to the one assumed in the diagram.

In a direct current network, only steady electric and magnetic fields are present. As we found in section 1.5.1, a steady electric field E is related to a scalar potential which is a function of position only. Around any closed path the potential returns to its starting value, i.e., the line integral of the field E is zero:



Kirchhoff’s voltage rule applies this equation to paths which follow the connections of the network. Along these paths potential differences occur only across the resistors and sources of e.m.f. As with the current rule, it is necessary to be careful about the signs of current and voltage. Around the loop ABDEA in Figure 8.1, for example, the e.m.f. Va balances the voltage drops across the resistors R1, R4 and R3, and



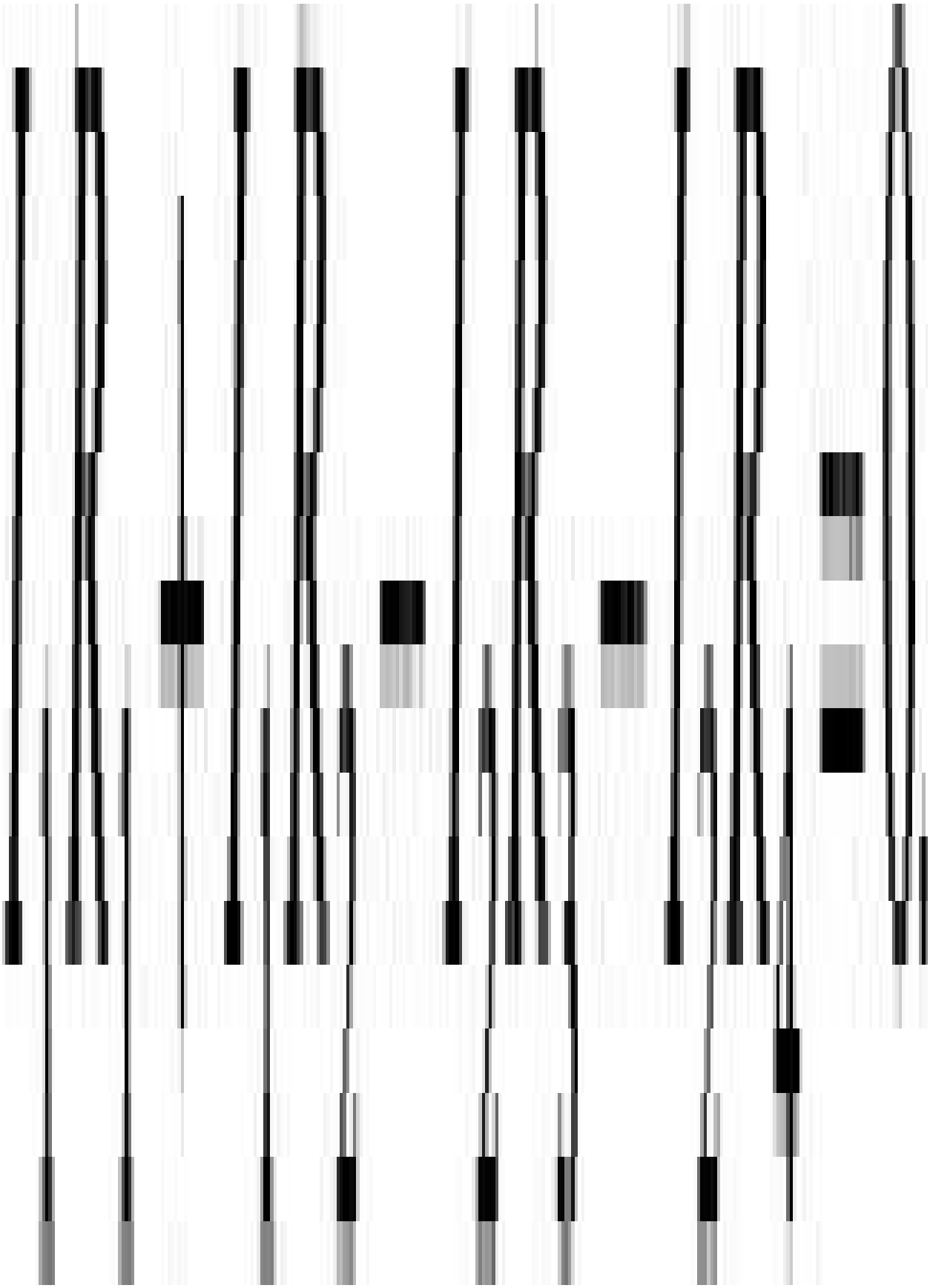
As an elementary example of network analysis, let us apply Kirchhoff’s rules to the Wheatstone bridge shown in Figure 8.2. The value of an unknown resistor R1 is to be determined by adjusting a variable standard resistance R2 until there is no potential difference between points B and D. Suppose that the detector used to search for the null point draws a current which is very small compared with the current flowing in the arms of the bridge even when it is off balance. Neglecting the current through the detector, Kirchhoff’s current rule tells us the same current I1 flows through the resistors R1 and R2, and similarly a current I2 through R3 and R4. To satisfy Kirchhoffs current rule at all nodes in the network, the battery must deliver a current (I1 + I2). Now apply the voltage rule to the loop ABCA, which passes through the battery;

(8.1)

and to the loop ADCA;

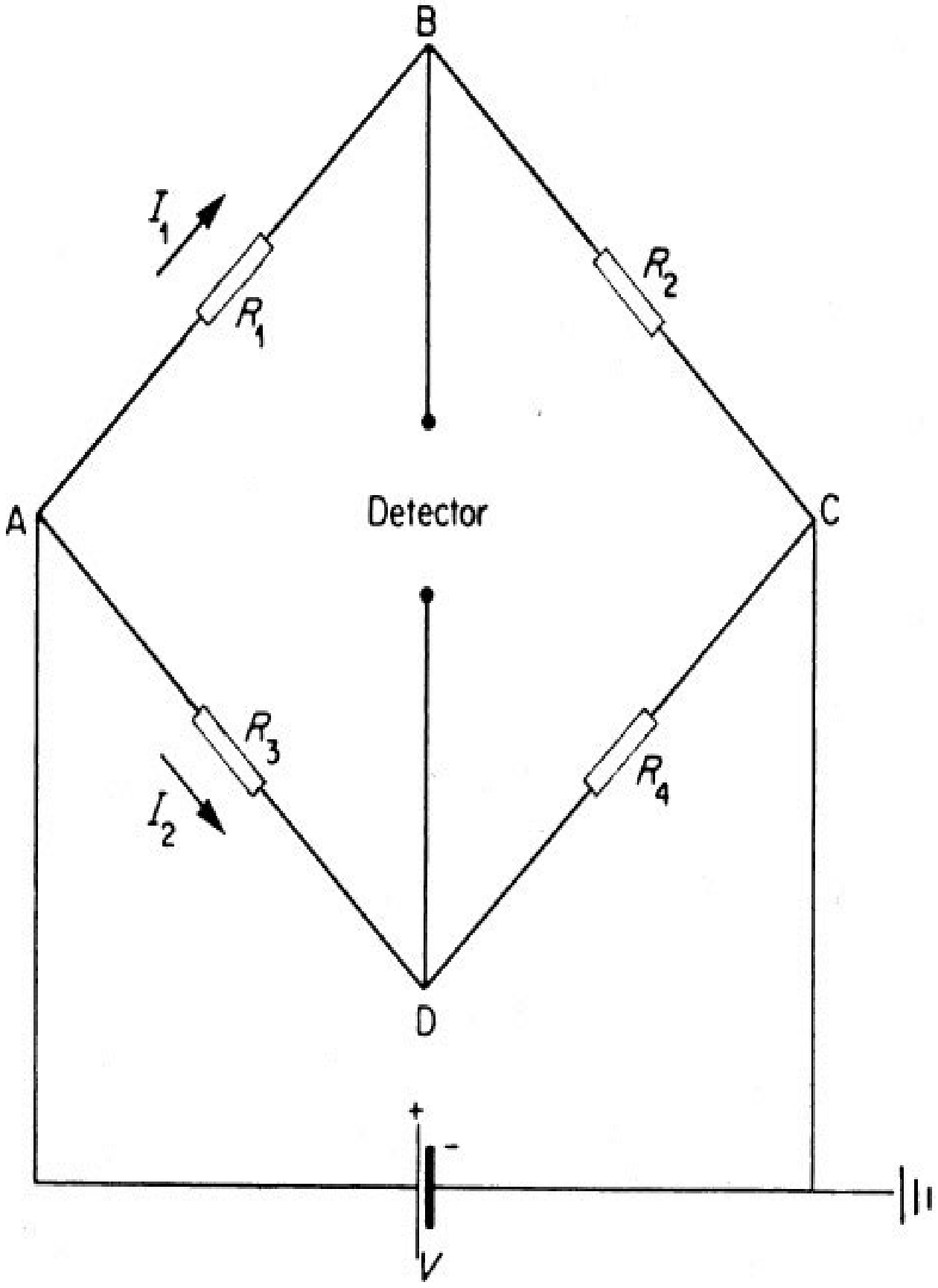
(8.2)

The only other loop which can be made along the current-carrying branches of the network is the loop ABCDA. The sum of the voltages around this loop is



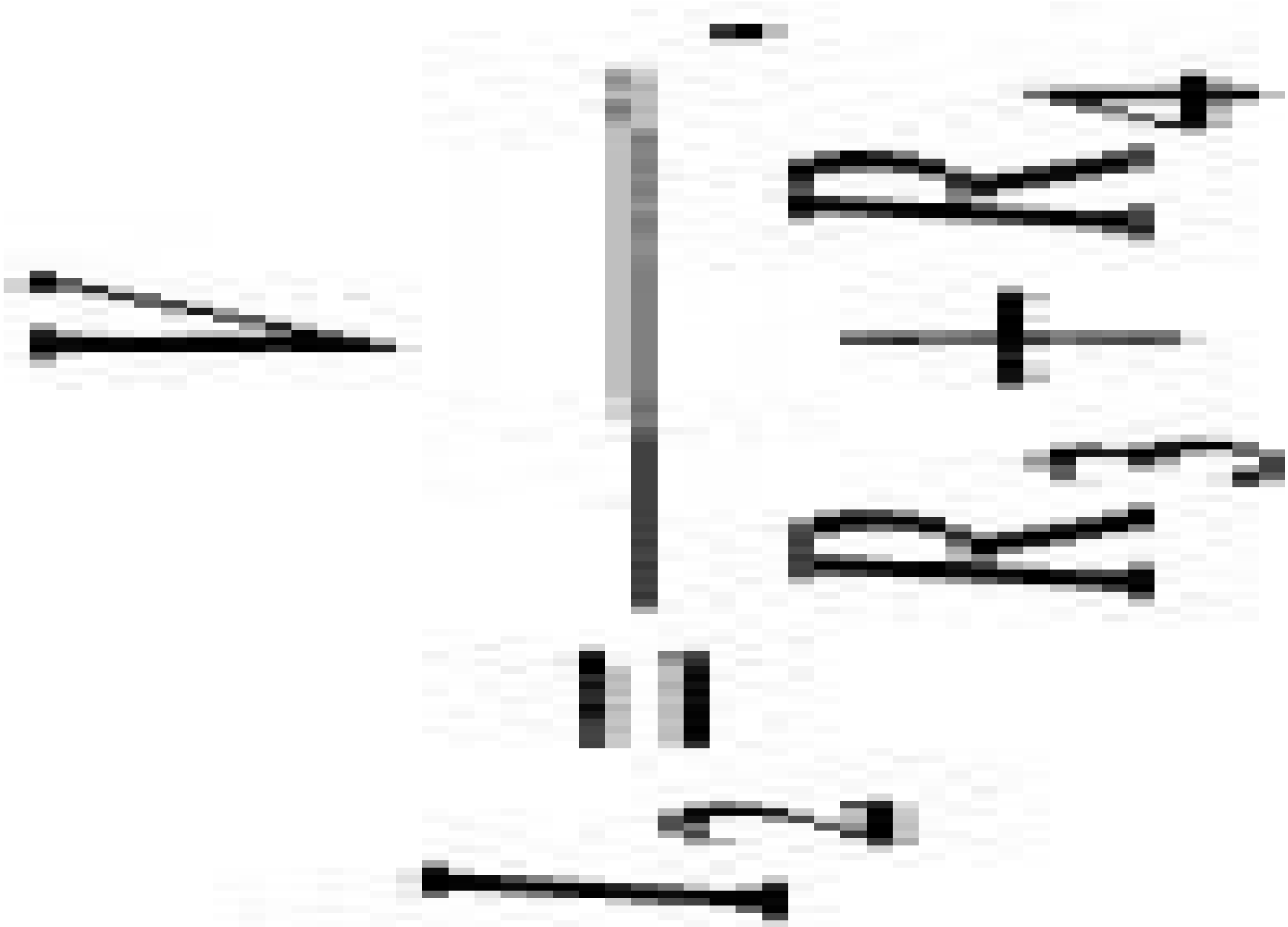
This equation can also be derived by subtracting Equation (8.2) from Equation (8.1). There are thus only two independent equations, from which solutions for the two unknowns I1 and I2 can be determined. For any network, Kirchhoff's rules always provide the same number of independent equations as there are unknown currents. When all the impedances in the network are linear, the set of linear simultaneous equations has a unique solution. Note, however, that if some of the circuit elements are non-linear, so that higher powers of the current appear in the equations derived from Kirchhoff's voltage rule, the solution may not be unique; it is quite possible for there to be more than one way of distributing currents in an electrical circuit with non-linear components.

Figure 8.2. The Wheatstone Bridge.


B
A
C
D
Detector
I_1
I_2
R_1
R_2
R_3
R_4
+
−
V

The Wheatstone bridge contains only resistors which are linear, and the currents are uniquely determined. It follows immediately from Equations (8.1) and (8.2) that

and



Hence the output voltage across the terminals of the detector is

(8.3)

and the null point occurs when v is zero, i.e. when

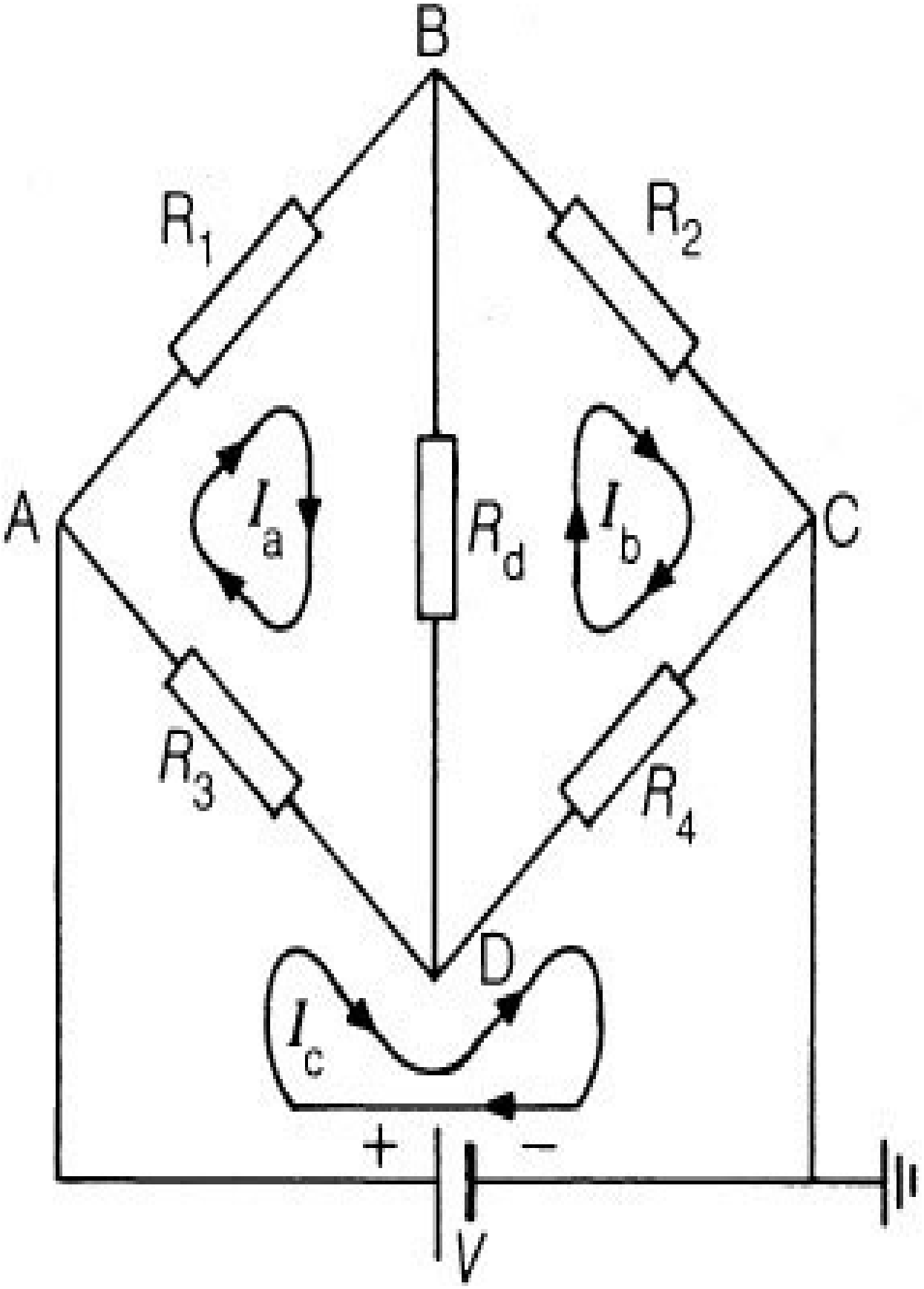
(8.4)

8.1.2 Loop analysis, node analysis and superposition

When solving the Wheatstone bridge problem, the currents I1 and I2 were chosen as the unknown variables. These were introduced as the currents flowing through the resistors R1 and R2. However, they can equally well be regarded as currents flowing round the loops ABCA and ADCA, which were the loops we used for the application of Kirchhoff's voltage rule. The branch CA is common to both these loops, and so the current through the battery is (I1 + I2), just the result found by applying Kirchhoff's current rule to node A or node C. If we choose always to consider loop currents, Kirchhoff's current rule is automatically satisfied, since as much current arrives at any node in a loop as leaves it. This is illustrated in Figure 8.3, which shows the Wheatstone bridge circuit again, with the addition of the resistor Rd to include the effects of current flowing through the detector. There are now three independent loop currents around the loops ABDA, BCDB and ADCA. We shall call the loop currents Ia, Ib and Ic as indicated in Figure 8.3. These currents clearly satisfy Kirchhoff's current rule at each node, and also determine the currents in all branches of the network. For example, the current through the resistor R3 is Ic – Ia. Application of the voltage rule to the same three loops then gives the three independent simultaneous equations needed to find Ia, Ib and Ic. This method of solving

network problems is called loop analysis.

Figure 8.3. A Wheatstone bridge circuit showing loop currents Ia, Ib and Ic.

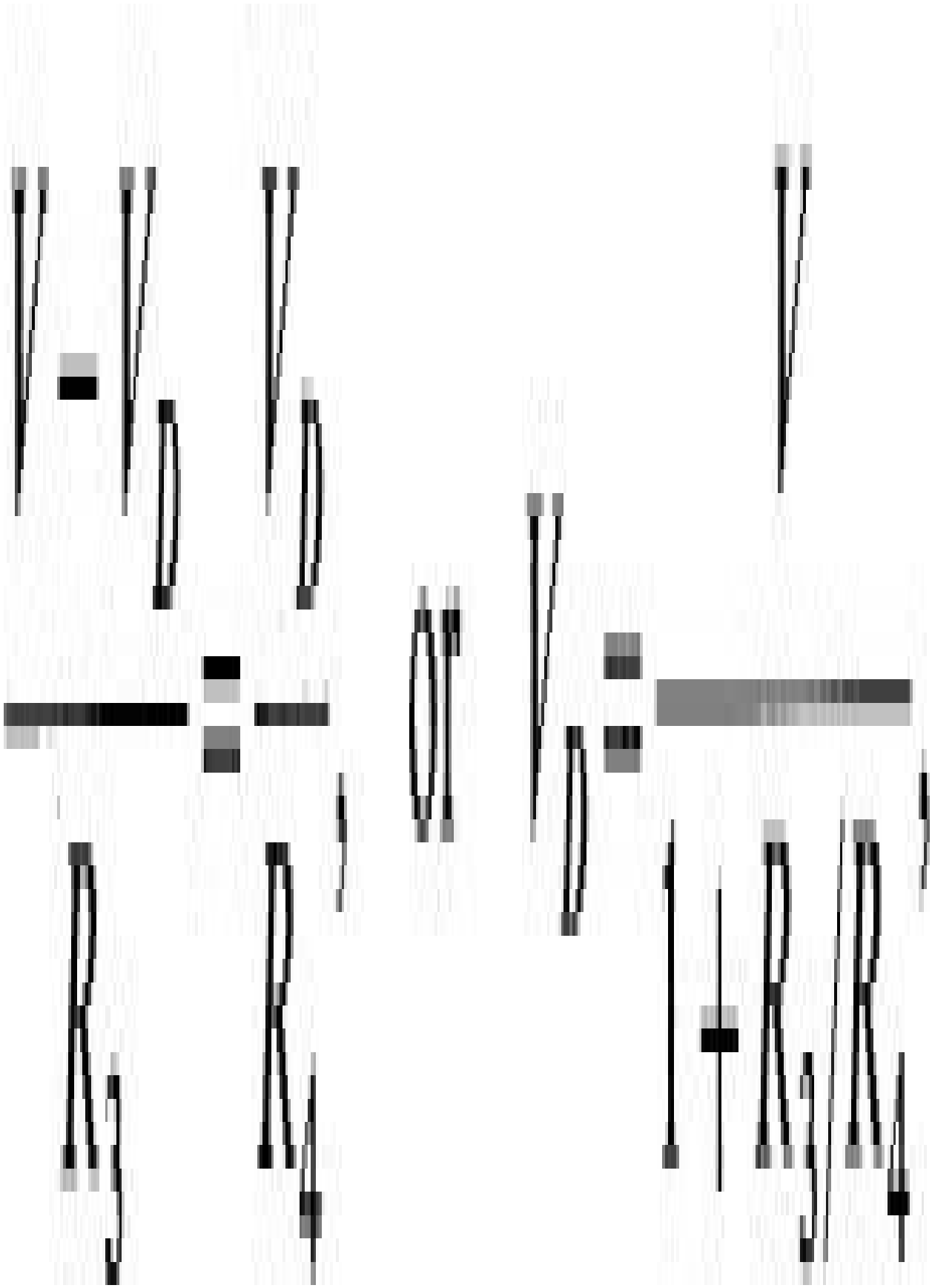

B
R1
R2
A
Ia
Rd
Ib
C
R3
R4
D
Ic
+
−
V

An alternative method is to choose node voltages as the unknown quantities, and to apply Kirchhoff's rules in the reverse order. This method is called node analysis. In the Wheatstone bridge, the voltages at A and C are known to be V and zero volts respectively. The unknown voltages are VB and VD, at the nodes B and D, and we choose these as the variables. In terms of these voltages Kirchhoff's voltage rule is automatically satisfied around any loop. Now we can apply the current rule to the nodes B and D, thus deriving the required simultaneous equations.

If we once again simplify the problem by assuming that no current flows through the detector, the currents in the branches AB and BC are the same, and

so that

Similarly at node D,

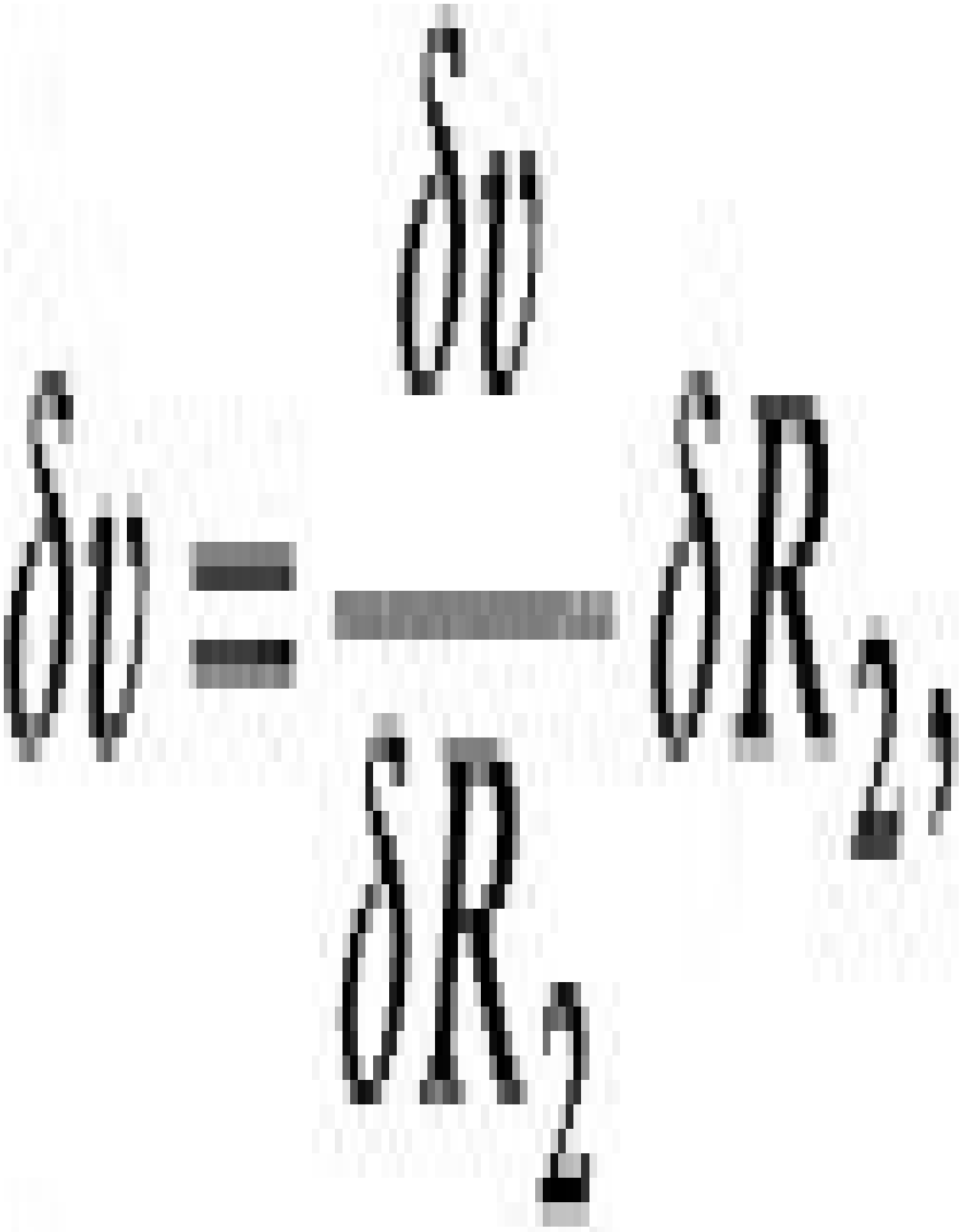


and

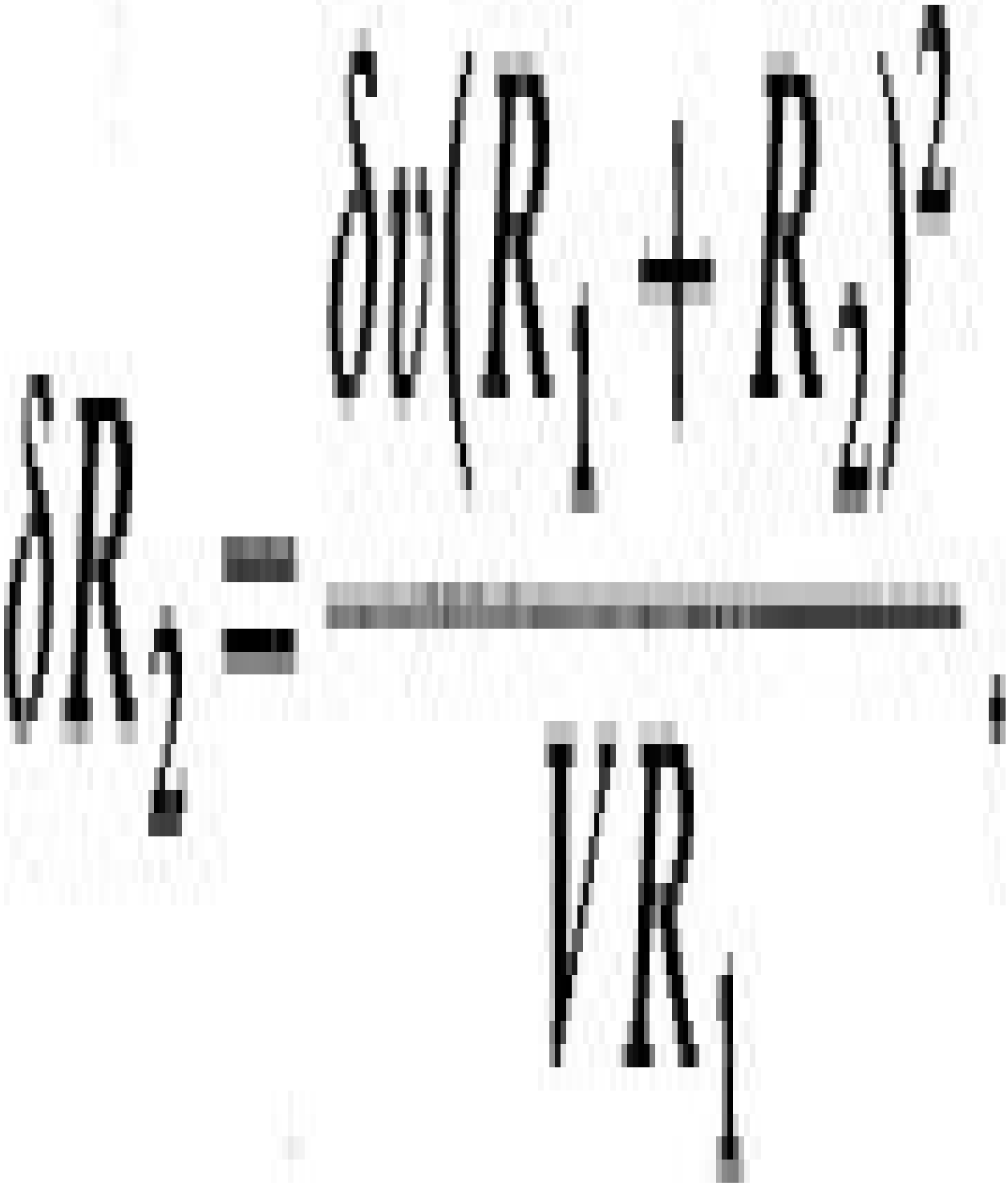
(8.5)

the same result as found in Equation (8.3) from the loop analysis. Notice that in this case no more variables are introduced if the detector has a resistance Rd which is not large compared with the other resistances in the circuit. There are still only two unknown voltages, at the nodes B and D, but the detector current (VB – VD)/Rd must now be accounted for when applying Kirchhoff's current rule at these nodes. For this problem the loop analysis has three unknown currents, and is less economical than the node analysis.

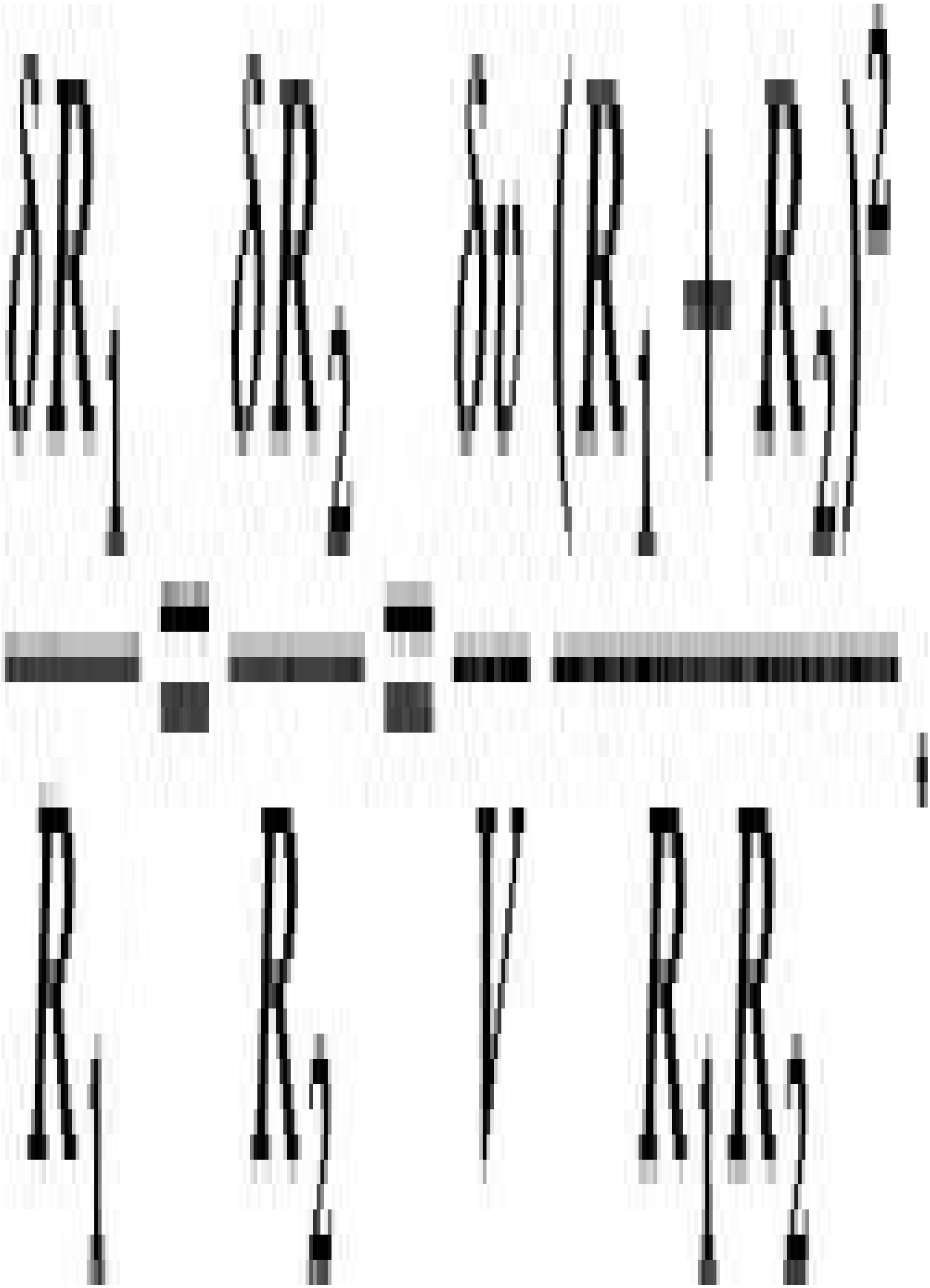
As well as giving the balance condition, Equation (8.5) tells us how sensitive the bridge is when used to measure R1. Suppose that a just detectable output voltage δv is caused by a change $\delta R2$ in the variable resistor from its balance position. We have



and by differentiating Equation (8.5) it follows that



Now the relative accuracy of measurement of the unknown is



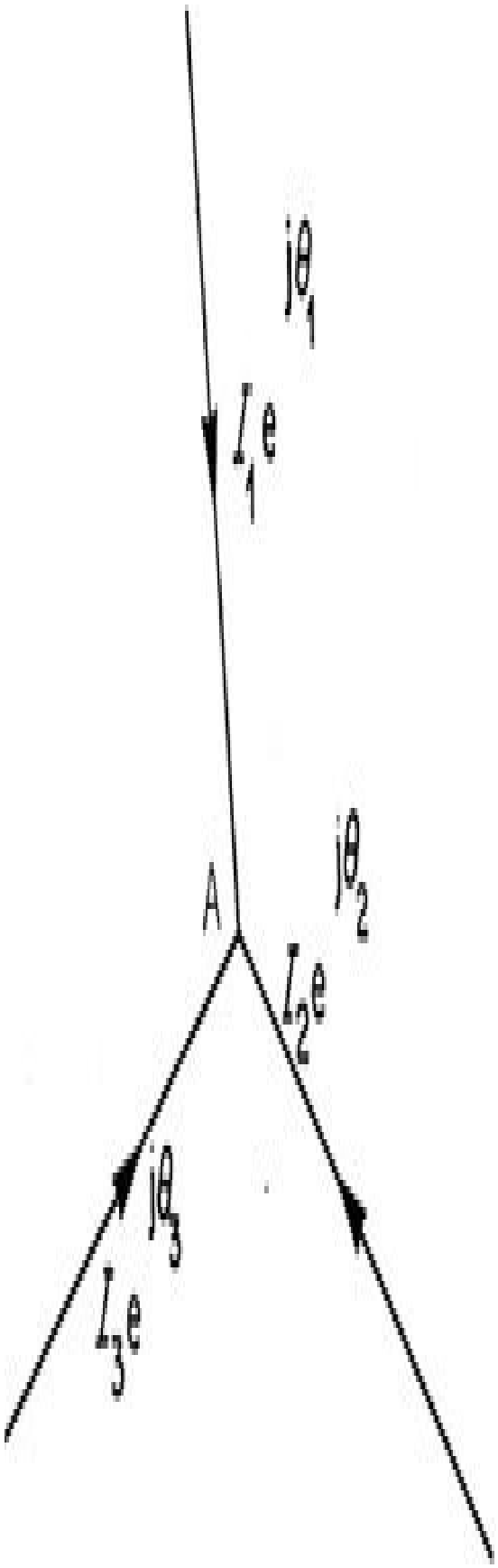
Since δv is fixed by the sensitivity of the detector, the least error in the unknown occurs when $(R1 + R2)^2/R1R2$ is minimum, i.e. when $R1 = R2$. At balance (Equation (8.4)) implies that $R3 = R4$, and the bridge is best used with equal values for these fixed resistors.

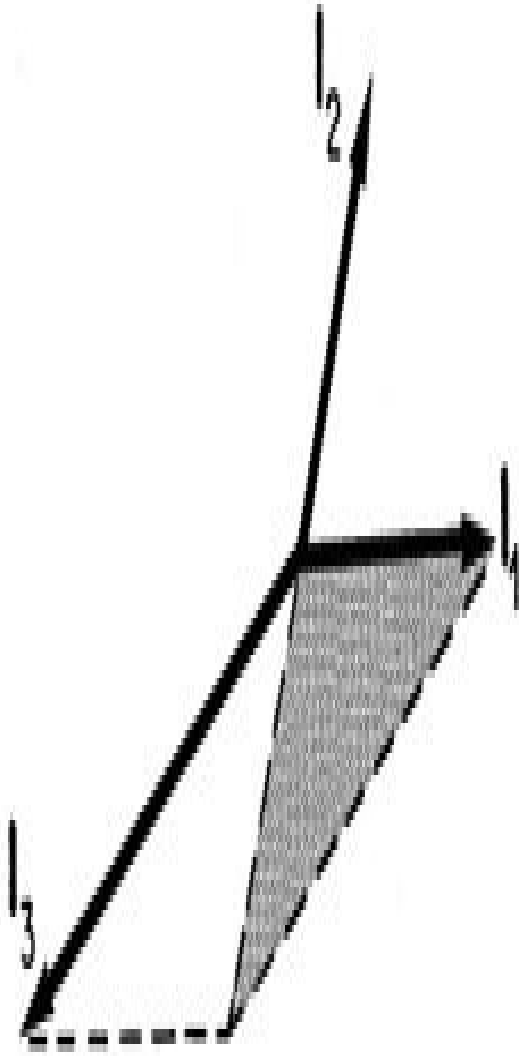
The practical limitation to the accuracy which can be attained in the measurement of resistance with the D.C. Wheatstone bridge is often not the size of the detectable output voltage, but the interference due to unwanted e.m.f.s in the circuit. The most important of these are usually thermo-electric e.m.f.s generated when the temperature of the circuit is not uniform. Because the circuit is linear, we can consider the effects of the thermo-electric e.m.f.s and of the battery separately. Imagine two networks representing a Wheatstone bridge, one containing no e.m.f. other than the battery, and the other only thermo-electric e.m.f.s. The current flowing in any branch of a network containing both types of e.m.f. is the sum of the currents in the same branch of the separate networks. This is an example of the application of the principle of superposition which was introduced in the discussion of electric fields in section 1.2: for linear networks containing more than one source of e.m.f., we simply superimpose the current distributions generated by each source on its own. We can therefore eliminate the effect of thermo-electric e.m.f.s by making a second measurement of R2 after reversing the battery. The unwanted e.m.f.s are not reversed at the same time, and provided that they are small, the mean of the two measurements of R1 is a good estimate of the true balance value. A still better way of eliminating the effect of thermo-electric currents by taking advantage of their independence is to replace the battery by an A.C. generator, and to use a detector sensitive only to signals of the same frequency as the generator. After generalizing Kirchhoffs rules to apply to A.C. networks, we shall show that the balance condition 8.4 still holds at low frequencies for the A.C. resistance bridge.

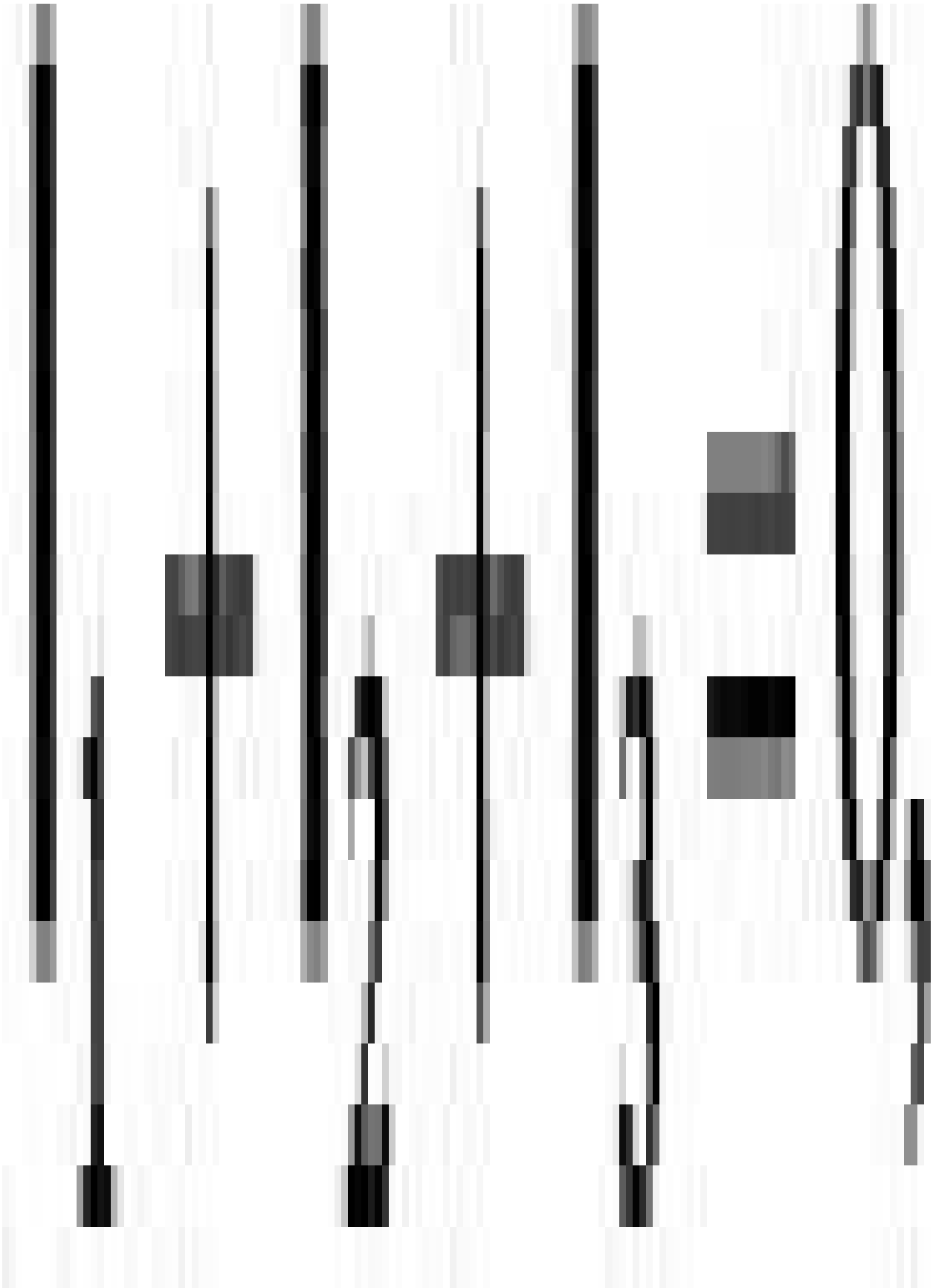
8.1.3 A.C. networks

Kirchhoff’s two rules apply for A.C. as well as D.C. In an A.C. network carrying currents of an angular frequency ω, conservation of charge still requires that at all times the current arriving at any node is zero. The currents in different sections of the network may have different amplitudes, and they may not be in phase with each other. Suppose, for example, that three leads carrying currents of amplitude I1, I2 and I3 meet at a node of the network. At time t = 0 their phase angles are θ1, θ2 and θ3, say, measuring the current in each lead in the direction towards the node. We may write Kirchhoff’s current rule in the same form for A.C. as for D.C. by using the complex number notation;

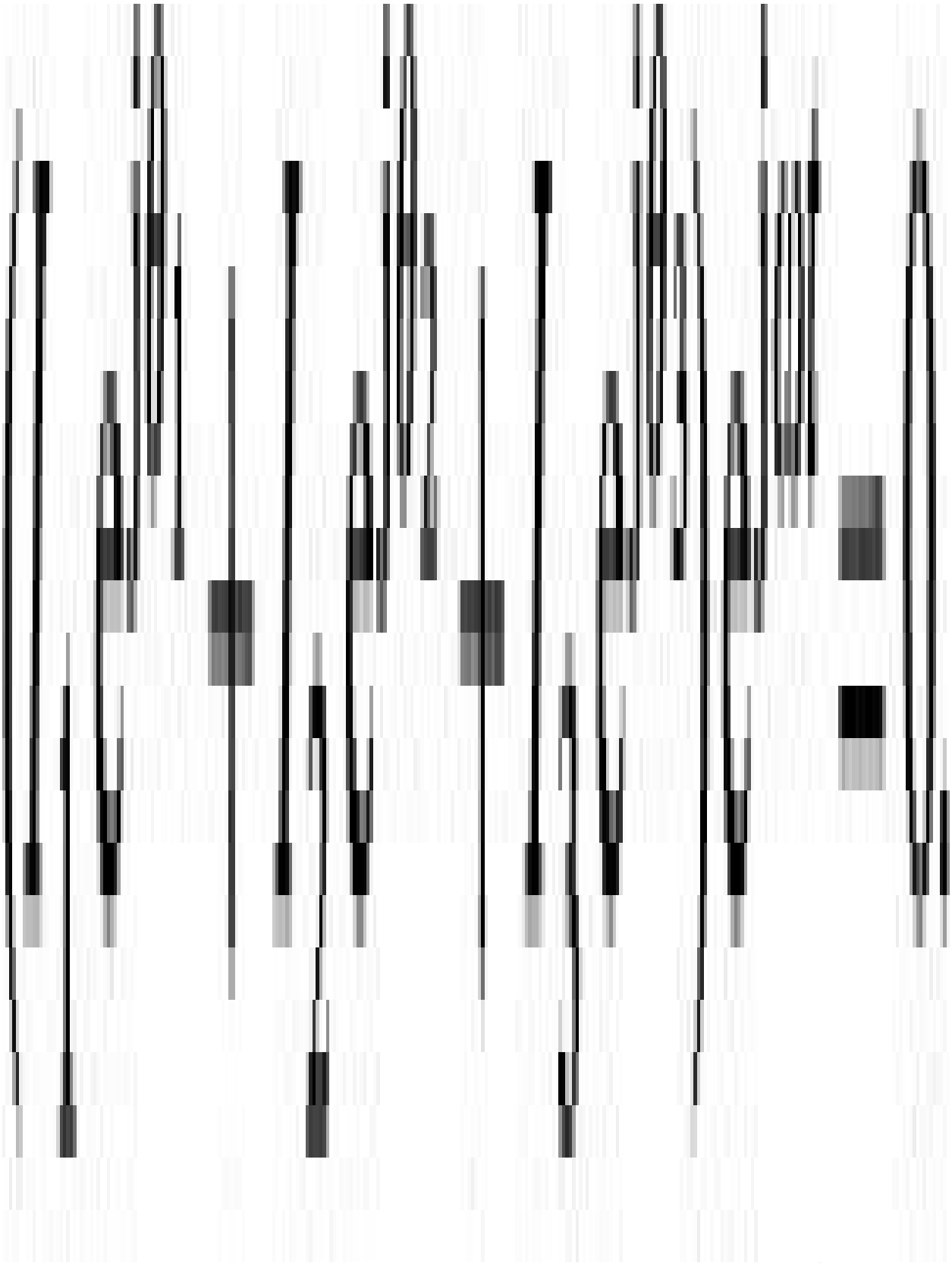
Figure 8.4. The bold arrows are phasors representing the currents arriving at the node A. To satisfy Kirchhoff’s current rule, the phasors must be able to form the sides of a closed polygon—in this case the shaded triangle in the figure.


jeθ1
I1e
A
jeθ2
I2e
jeθ3
I3e


I2
I1
I3



or in full



The real part of this equation requires that the actual currents satisfy Kirchhoff's current rule at all times. The imaginary part has no physical meaning, but allows us the convenience of using complex numbers in solving network equations. Figure 8.4 shows a phasor representation of three currents arriving at a node: to satisfy Kirchhoff's rule, both the x- and y-projections of the phasors must sum to zero, which requires that they can be aligned along the sides of a closed polygon.

The changing currents in an A.C. network cause changing magnetic fields, and these lead to induced e.m.f.s. Faraday's law tells us that the e.m.f. around a closed path is $E \cdot dl = -d\Phi/dt$, where Φ is the total magnetic flux threading the area enclosed by the path. Kirchhoff's voltage rule for an A.C. network applies Faraday's law to a path following the connections of the network. As for D.C., the rule states that the sum of the voltages around the loop is zero: note, however, that the sum must now include the voltage drops due to induced e.m.f.s. Induced e.m.f.s are generated by the self-inductances and mutual inductances of coils, but in addition there are always 'stray' inductances associated with the leads between the components; even in an A.C. network containing no coils, changing fluxes thread the areas enclosed within the network. At moderately low frequencies the stray inductances are very small and they can usually be neglected. This is fortunate, since their magnitudes depend on the exact shape and dimensions of the circuit, and are generally almost impossible to calculate.

Not only are there always small inductances present in an A.C. network, but there is inevitably stray capacitance between different parts of the circuit, and between different parts of the same component. We have already met stray capacitance in the discussion in section 7.4 of the current through an inductive coil, which has capacitance between neighbouring windings. Neglecting this capacitance, the coil can be represented by an approximate equivalent circuit consisting of a resistance R in series with an idealized self-inductance L. Similarly, by neglecting all the stray inductances and capacitances which have little effect at low frequencies, an equivalent circuit can be built up for the whole of an A.C. network. Each component is represented by an idealized circuit

element, and the idealized elements are linked together in the same way as the components of the real circuit. The leads are assumed to have no impedance, and voltage differences only occur across the idealized elements. Just as for a D.C. network, application of Kirchhoff's two rules to the equivalent A.C. network leads to a set of linear equations whose solution yields the currents in all branches of the network.

Let us estimate how good an approximation it is to neglect stray impedances in the Wheatstone bridge circuit of Figure 8.2 if the battery is replaced by an A.C. generator. Let the resistance in each arm be 1000 Ω. The self-inductance of a circular loop of radius 10 cm made of 1mm diameter wire is 0.7 μH. The dimensions of the wires of a typical bridge are of this order of magnitude, and hence the stray inductance in the bridge will also be about 1 μH. At a frequency of 1 MHz the magnitude of the impedance of a 1 μH inductance is ωL 6 Ω, and the induced e.m.f. is less than 1% of the voltage drop across the resistors. It was shown in section 3.3.1 that the capacitance between a pair of parallel wires separated by three times their diameter is 15 pF/m. Stray capacitance in ordinary circuits therefore amounts to a few picofarads. At 1 MHz the magnitude of the impedance of a 1 pF capacitance is l/ωC 2.10^5 Ω. Such a large impedance can shunt less than 1% of the current carried by the resistance.

No hard and fast rules can be given for the frequency limit above which the neglect of stray capacitance and inductance ceases to be a good approximation. The example worked out above suggests that to achieve an accuracy of 1% one may have to take strays into account at frequencies above 1 MHz. To some extent this can be done by adding to the equivalent circuit discrete capacitances and inductances representing the strays. It is difficult to find accurate values for the impedances of these extra circuit elements, but even a rough guide to the high frequency behaviour of a circuit may be useful. Eventually the network analysis must break down at frequencies above about 10^9Hz. Stray inductances and capacitances cannot both be assumed to cause small corrections at these frequencies; any component is either shunted by stray capacitance or causes a smaller voltage drop than the e.m.f. generated by stray inductance. However, at low frequencies, it is usually safe to neglect stray impedances in a Wheatstone bridge with resistors in each arm. If a detector which responds only to alternating

voltages is used, unwanted thermo-electric e.m.f.s are not observed, and the null-point occurs when the balance condition (8.4) is satisfied.

8.2 AUDIO-FREQUENCY BRIDGES

The Wheatstone bridge circuit is used at audio-frequencies (i.e. at frequencies up to about 20 kHz) to measure the impedance of reactive elements. The equivalent circuit for the general A.C. bridge is shown in Figure 8.5. The arms of the bridge contain circuit elements with complex impedances Z1, Z2, Z3, Z4. The currents I1 and I2 may not be in phase with the generator voltage or with each other. In order to obtain a balance with no alternating voltage across the detector, the voltages across Z1 and Z3 must not only have the same amplitude, but they must also be in phase with one another. The balance condition is found by applying Kirchhoff's second rule to the bridge. The resulting equations have the same form as they did for the D.C. bridge, with the complex impedances replacing resistances. The balance condition becomes

(8.6)

This equation has real and imaginary parts, thus imposing the two conditions on the component values required to ensure that both the amplitude and the phase of the voltages across Z1 and Z3 are the same. An A.C. bridge must contain two variable components, and in searching for a balance these are adjusted alternately to give successive minima in the detector output until the null-point is reached.

There are many forms of the A.C. bridge, with different combinations of components in the arms. As a practical example we shall consider the Owen inductance bridge, which has the circuit diagram shown in Figure 8.6. The self-inductance L and the resistance R represent a coil whose inductance is to be measured. The bridge is balanced by adjusting the variable resistors R1 and R3. Substituting the appropriate values in Equation (8.6), at balance we have

$$\frac{j\omega L + R + R_1}{R_2} = \frac{R_3 + 1/j\omega C_3}{1/j\omega C_4}$$

$$= \frac{C_4}{C_3} + j\omega C_4 R_3,$$

Figure 8.5. The alternating current Wheatstone Bridge.

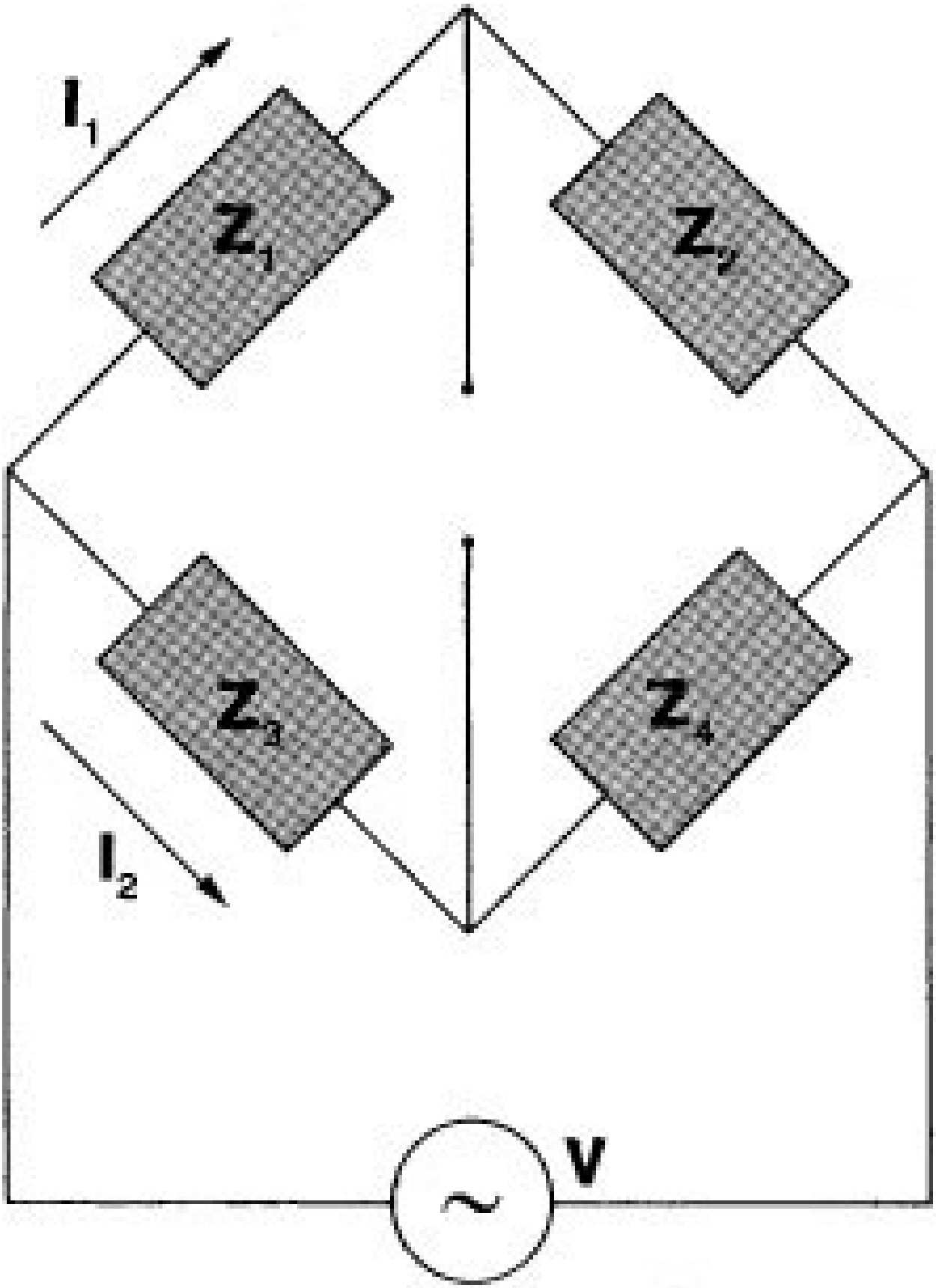
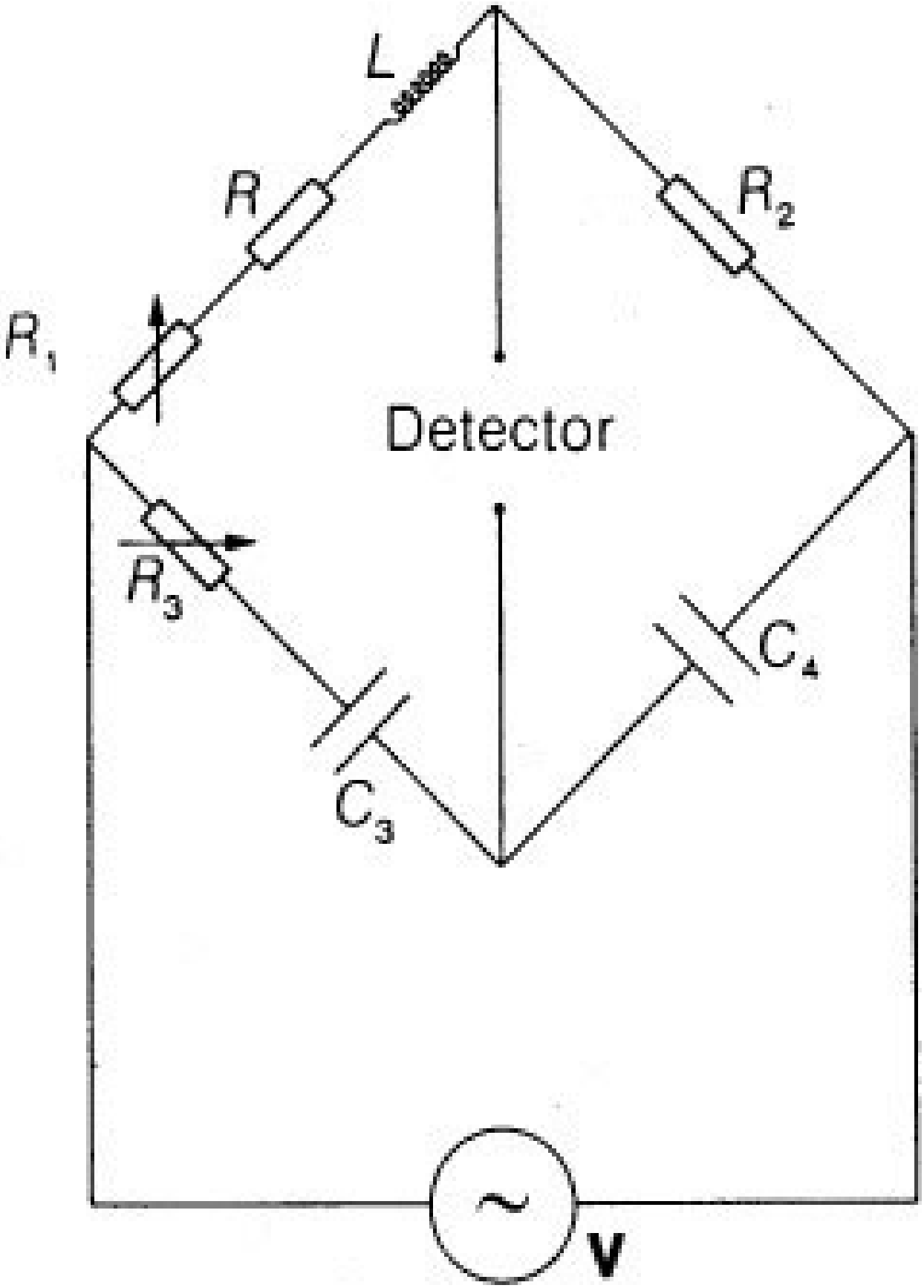

I1
Z1
Z2
Z3
Z4
I2
V

Figure 8.6. The Owen inductance bridge.


L
R
R2
R1
Detector
R3
C4
C3
~
V

Equating real parts

(8.7)

and imaginary parts,

(8.8)

The Owen bridge has the following good features which make it convenient for routine laboratory use:

(i) The balance condition is independent of frequency. The accuracy of measurement is not affected if the time variation of the generator voltage is not exactly sinusoidal, and other angular frequencies besides ω are present in the circuit.

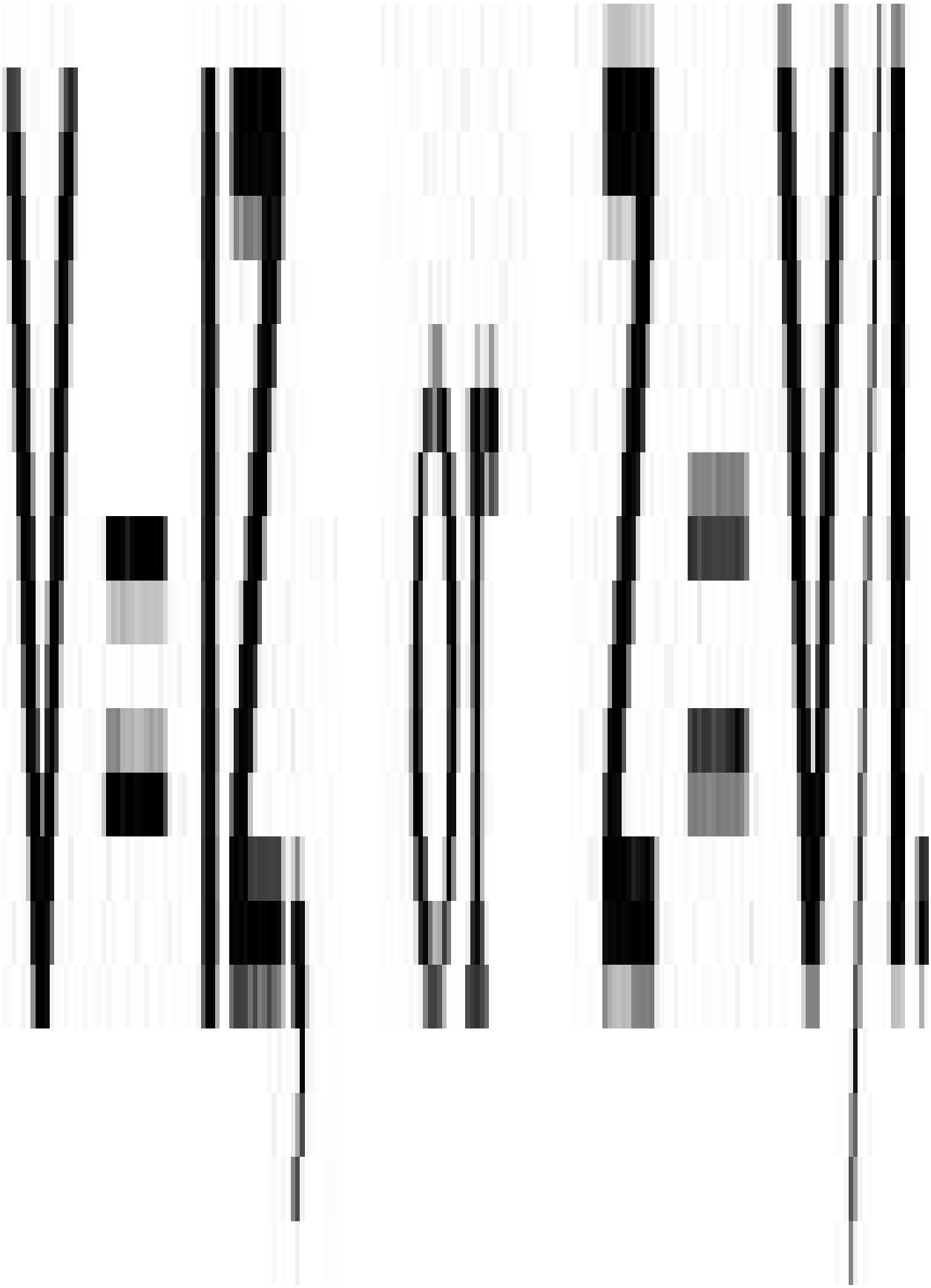
(ii) The resistance R1 does not occur in Equation (8.8). The balance setting of R3 is fixed by the inductance L of the coil, whatever its resistance R. It follows that the resistance R3 can be calibrated to give a direct reading of inductance. In fact the inductance is proportional to R3, with the proportionality constant known from the values of the fixed components C4 and R2. Similarly, since R3 does not occurs in Equation (8.7), R1 can give a direct reading of R.

(iii) The bridge is suitable for measuring small inductances. If the inductance to be measured is much less than 100 μH, it is no longer permissible to neglect the stray inductances in components and leads. Their effect can be eliminated by repeating the measurement after removing the coil and joining together the leads which were attached to it. If the new value of the variable resistance at balance is

R′3, the strays are equivalent to an inductance C4R2R′3 in the same arm of the bridge as R1. Evidently the inductance of the coil is C4R2(R3 – R′3).

8.3 IMPEDANCE AND ADMITTANCE

At a specified angular frequency ω there is a well-defined value for the impedance between any pair of nodes in a network consisting of linear passive elements. To find this impedance, one imagines that a source of e.m.f. generating a voltage $V = V0 \exp(j\omega t)$ is connected across the two nodes. By applying Kirchhoff's rules the currents in all the branches of the network can be calculated, and in particular the current I delivered by the generator. Since the circuit is linear, I is proportional to V, and we can write

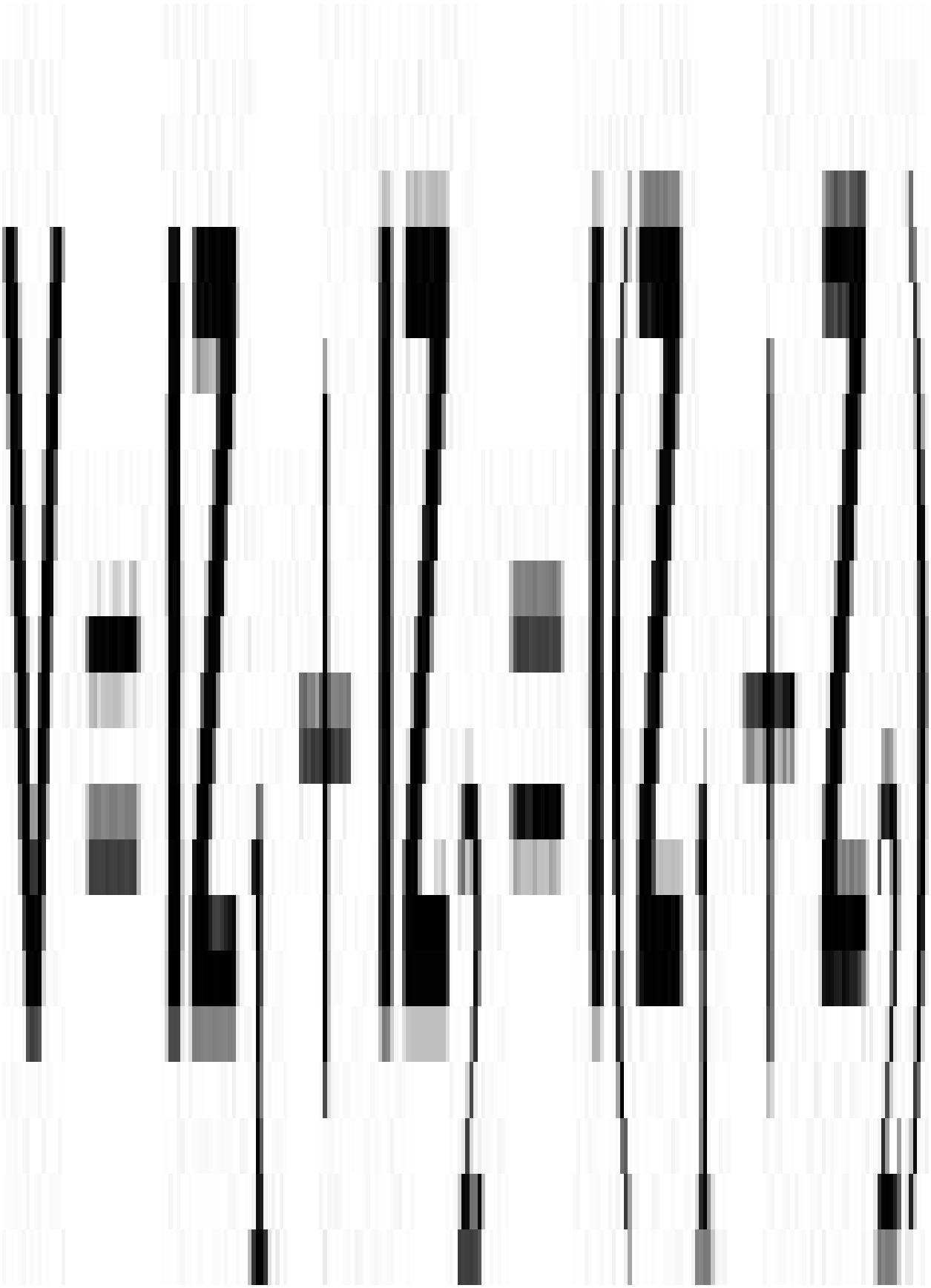


The quantity Z is the required impedance of the whole network between the two nodes connected to the terminals of the generator.

In general the impedance Z can only be found by solving a set of simultaneous equations derived by the application of Kirchhoff's rules. For elements connected in series or in parallel, these equations are not coupled, and expressions for the impedance of the network can be written down at once.

(a) Series networks

The two components with impedances Z1 and Z2 in Figure 8.7 are connected in series, with no node between them leading to other components. If a generator of voltage V is connected across the free ends of the components, then by Kirchhoff's first rule, the same current I flows through Z1 and Z2. The voltage drop across them is



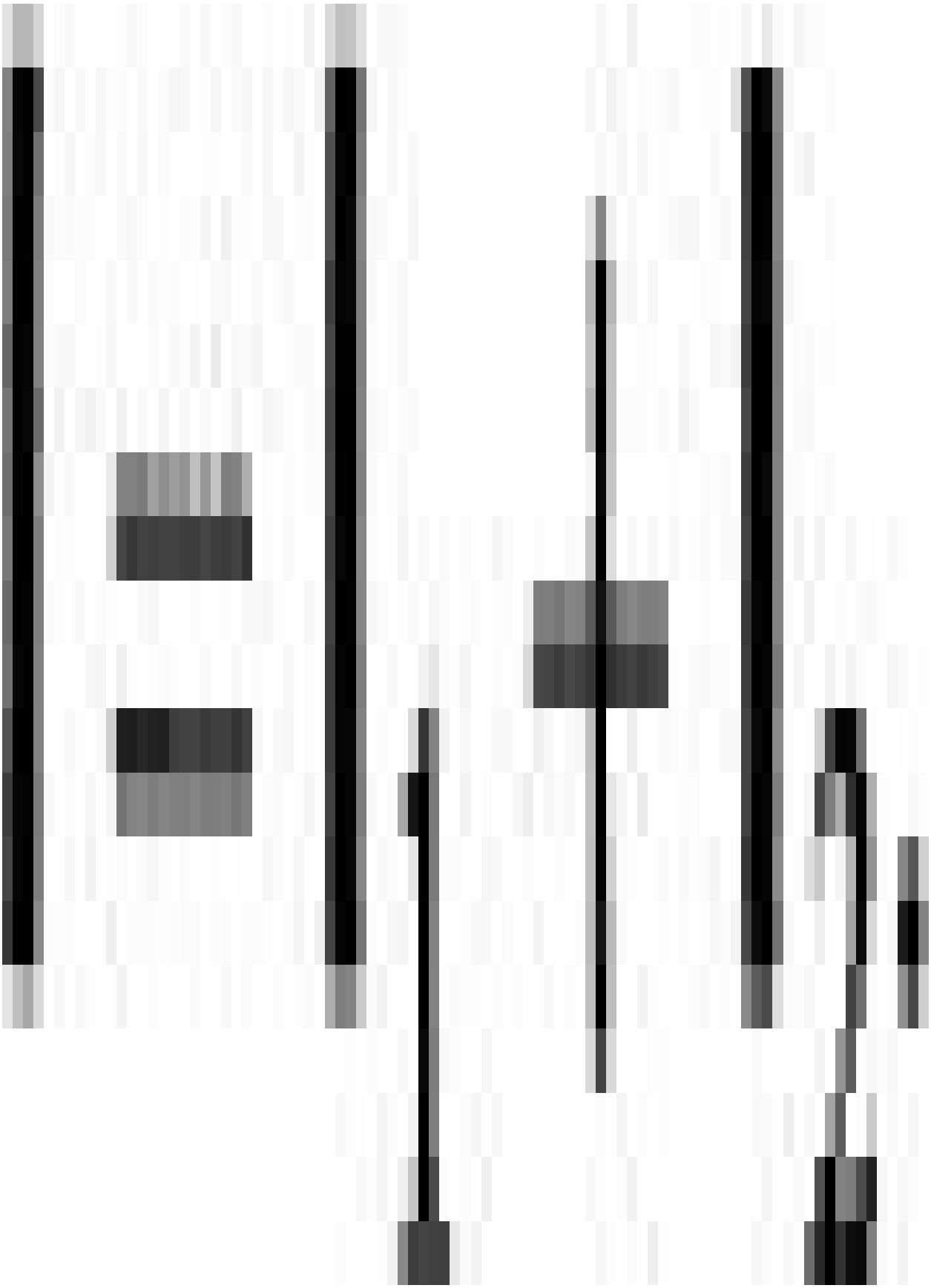
Then impedance of the pair of components is

(8.9)

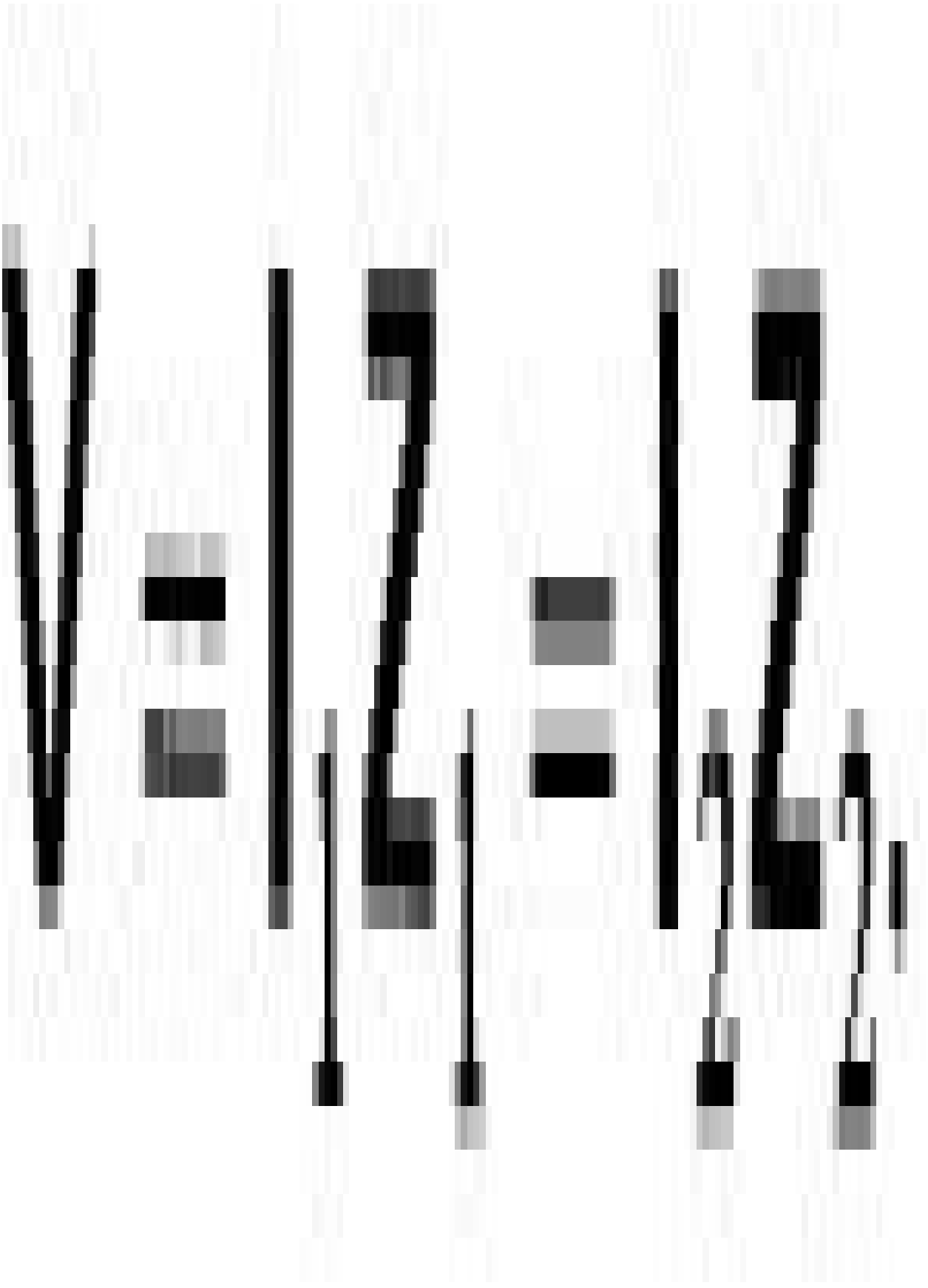
Similarly if any number of components are connected in series, the impedance between the ends of the series network is the sum of the impedances of each component separately.

(b) Parallel networks

When the parallel network of Figure 8.8 is connected across a generator of voltage V, currents I1 and I2 flow through the components Z1 and Z2 respectively, and the current delivered by the generator is



The whole of the generator voltage is applied across each component, so that



Hence

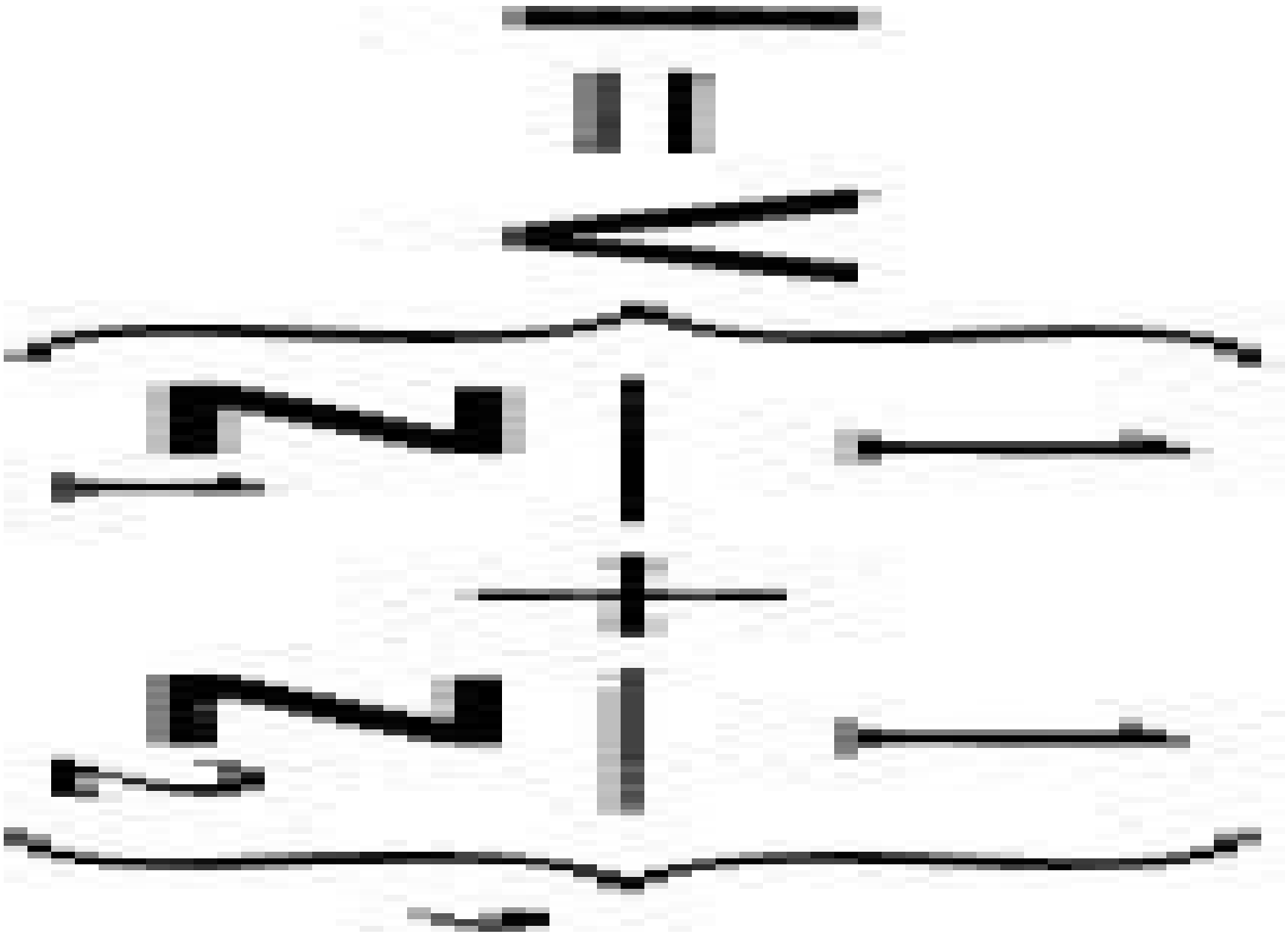


Figure 8.7. Impedances connected in series.

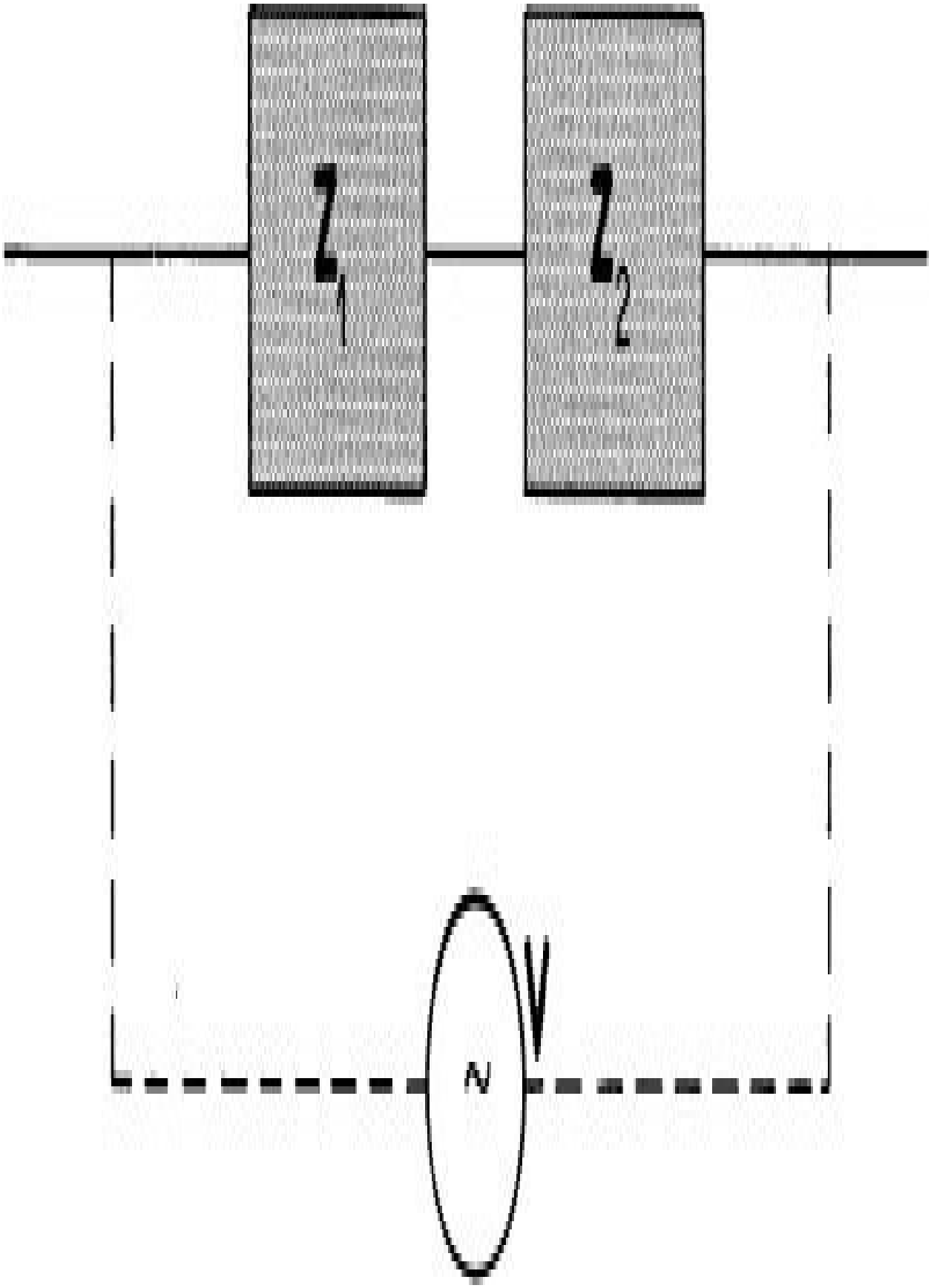
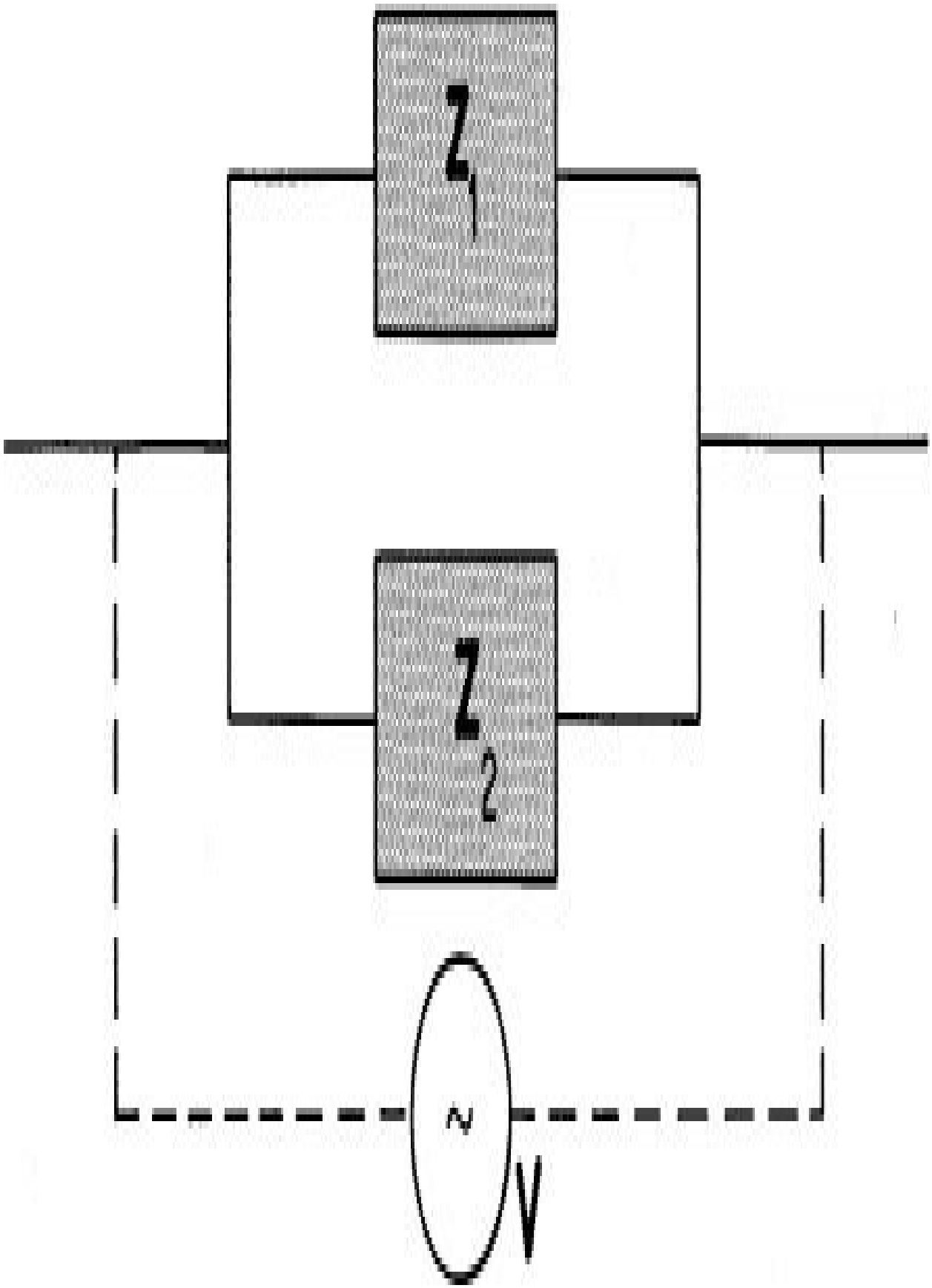

Z1
Z2
N
W

Figure 8.8. Impedances connected in parallel.


Z1
Z2
N
V

and

(8.10)

The reciprocal of impedance is called admittance, and is denoted by the symbol Y. For the parallel network

(8.11)

Similarly, if any number of components are connected in parallel, the admittance of the network is the sum of the admittances of each component separately.

The imaginary generators, shown connected by dotted lines in Figures 8.7 and 8.8, were introduced in order to drive current through the series and parallel networks, so that we could work out the current through each component and the voltage drop across it. But the impedances derived in Equations (8.9) and (8.10) do not depend on the voltage of the generators; these equations represent the impedances of the networks whether or not the generators are there. Let us suppose that the impedance between two nodes of some network is Z, and that these two nodes are connected to other components to form part of a larger network. Provided that the original network has external connections only at these two nodes, it can be replaced by a single component of impedance Z for the purpose of working out the currents in the rest of the network.

When a network consists of a combination of series and parallel connections, its total impedance can easily be found by successively calculating the impedance

of larger and larger sections. As an example we shall calculate the impedance at an angular frequency ω between the points A and D of the network in Figure 8.9. First we sum the impedances R2 and jωL in the series section BCD and replace them with the single impedance (R2 + jωL). This impedance is in parallel with the capacitor, and the total admittance Y between B and D is the sum

$$Y = \frac{1}{\frac{1}{R_2 + j\omega L}} + j\omega C$$

$$= \frac{1 - \omega^2 LC + j\omega C R_2}{R_2 + j\omega L}.$$

Figure 8.9. A network which is made up entirely of series and parallel connections.

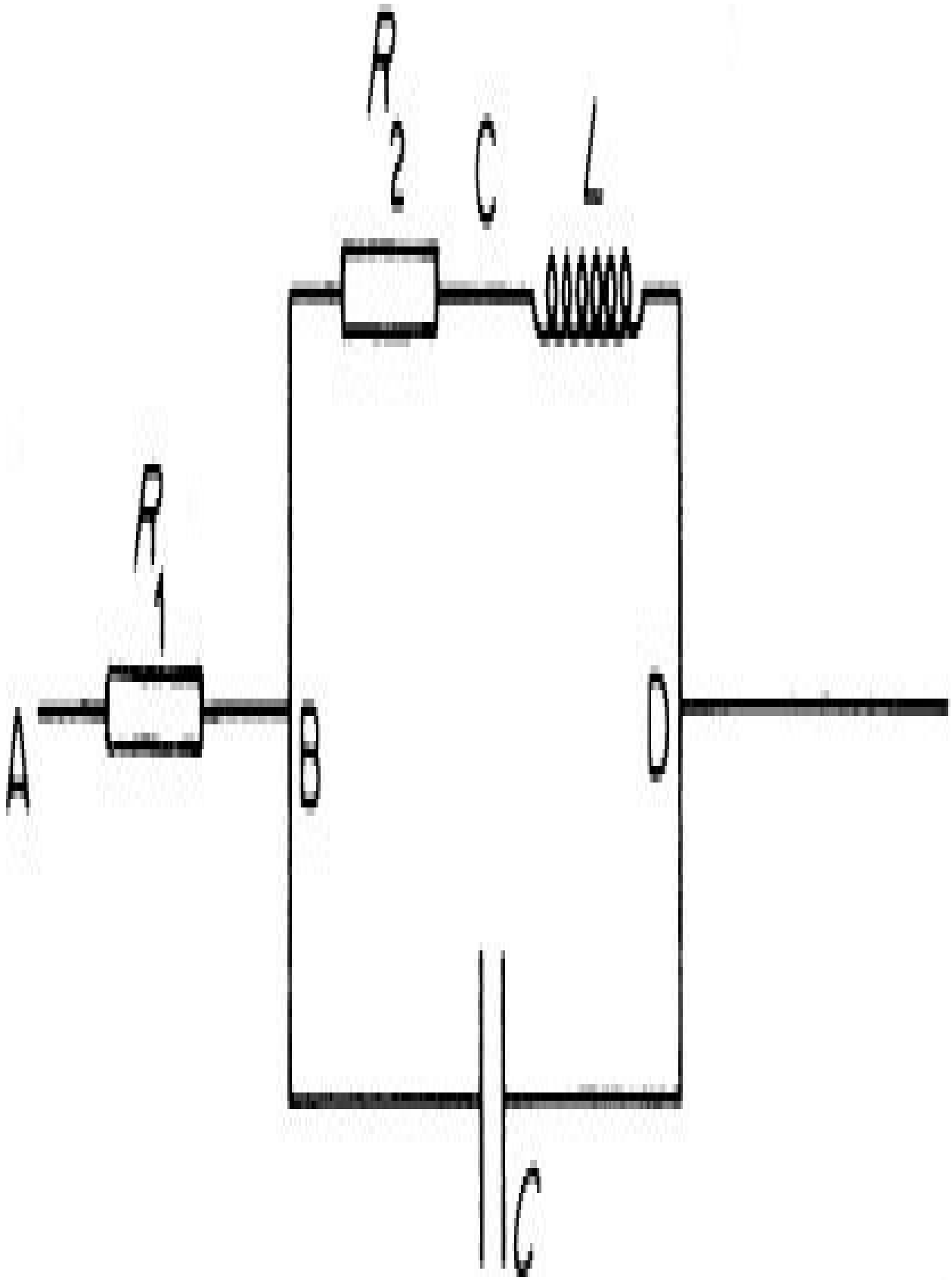
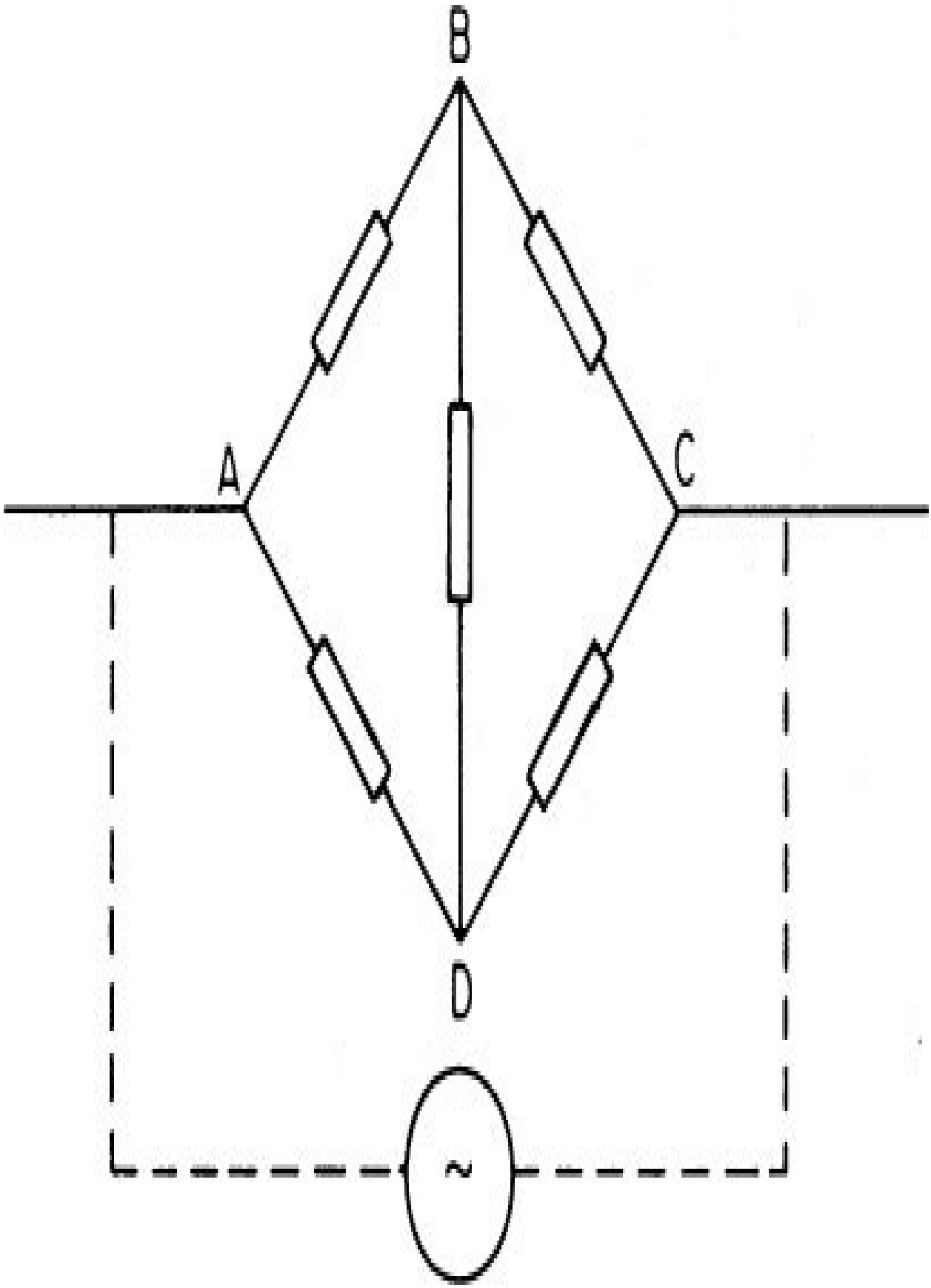

R2
C
L
R1
A
B
D
C

Figure 8.10. A network which cannot be broken up into sections connected in series or in parallel.


B
A
C
D
~

The impedance Y^{-1} of this section is in series with the resistor R1 and finally the impedance of the whole network between A and D is

$$Z = R_1 + Y^{-1}$$

$$= R_1 + \frac{R_2 + j\omega L}{1 - \omega^2 LC + j\omega C R_2}$$

Notice that if ωCR2 is small the denominator of the second term becomes small when $\omega^2 = 1/LC$. The network is resonant, and it has a high impedance near the resonant frequency = 1/LC of the series LCR circuit considered in section 7.6.

Not all networks can be broken down into series and parallel combinations. The simplest network which will not reduce in this way is the bridge network in Figure 8.10. If we wish to find the impedance between A and C we must connect an imaginary generator and apply Kirchhoff's rules to the whole network, since there are no subsections which are connected to the rest of the network at only two points.

8.3.1 Input impedance

The idea of splitting up electrical circuits into sections is an important one. Most complicated circuits are in fact designed in sections which have separate functions and are more or less independent of the another. In a radio set, for example, the tuner which selects the broadcasting station, the amplifier which drives the loudspeaker and the loudspeaker itself are different sections. Indeed the loudspeaker is often mounted in a separate cabinet, joined to the amplifier only by a lead carrying the signal and a common earth. The impedance of the loudspeaker between the signal lead and earth is called its input impedance. If the loudspeaker is a linear circuit, and it draws a current Iinput when connected to a signal voltage Vinput, then the input impedance is

(8.12)

For the purpose of analysing the behaviour of a network connected to the input of the loudspeaker, the loudspeaker may be represented by a single element with

impedance Zinput. Usually an input impedance can also be defined for a circuit which contains active and non-linear components. For example, if an audiofrequency amplifier is not to distort a signal, its response must be linear at the signal frequency. Throughout the amplifier, changes in current at this frequency are proportional to the signal voltage, even if they are superimposed on nonlinear steady currents. The effective input impedance for the signal is

(8.13)

where Vinput and Iinput are the voltage and current at the input lead. For linear circuits, this generalized definition of input impedance reduces to Equation (8.12).

8.3.2 Output impedance and Thévenin's theorem

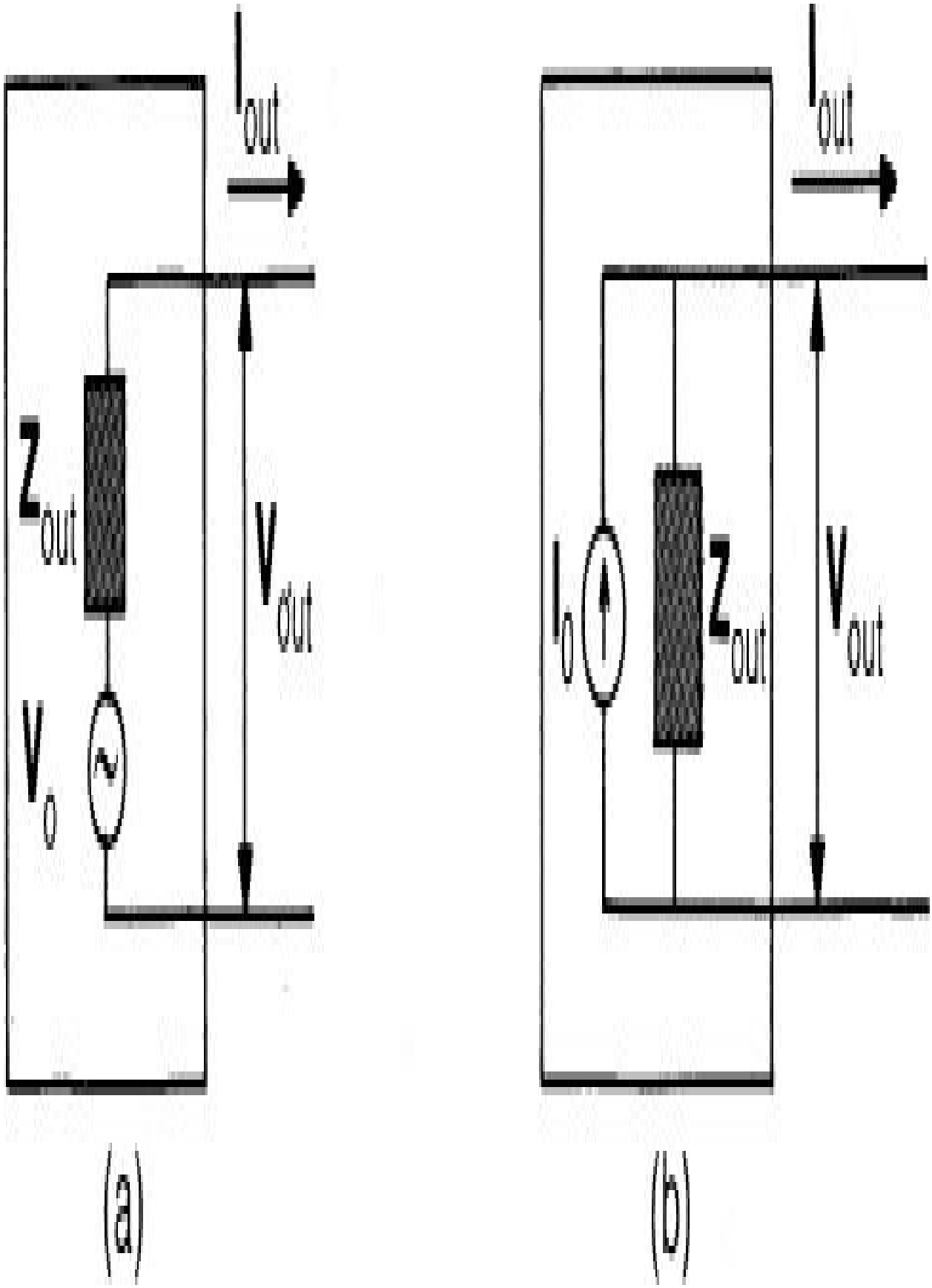
Similar considerations apply to the discussion of output signals, though now we must bear in mind that when there is an output signal, there must somewhere be a source of e.m.f. Suppose an electric circuit is in a box, and that it generates a signal across a pair of output terminals. The voltage Vout actually appearing across the terminals depends not only on the circuit generating the output, but also on the network to which the terminals are connected. If a high current is drawn from the output then the voltage will fall. For a linear circuit, if the current is Iout we may write

(8.14)

where V0 and Zout are constants.

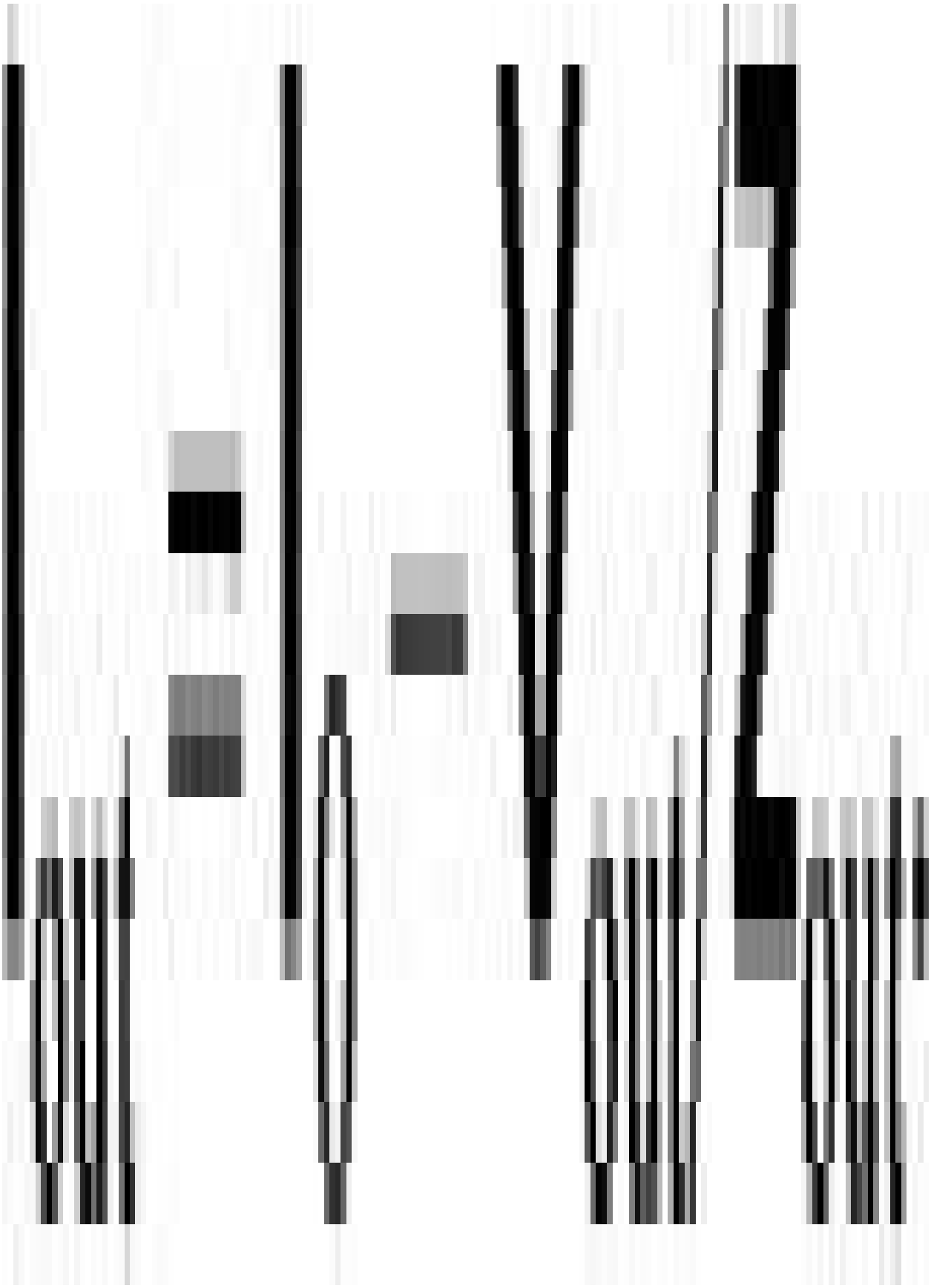
Thévenin's theorem states that when the circuit is connected to an external network, it may be represented by a single e.m.f. V0 in series with an impedance Zout. This is immediately apparent from Figure 8.11(a), since the network in the figure satisfies Equation (8.14) just as does the real circuit. This network is referred to as the Thévenin equivalent circuit. The impedance Zout is the output impedance of the circuit: if the e.m.f. in the Thévenin equivalent circuit is reduced to zero (i.e. replaced by a short circuit), then the impedance seen between the terminals from outside the box is Zout.

Figure 8.11. (a) The Thévenin equivalent circuit and (b) the Norton equivalent circuit.



(a) (b)

Sometimes it is more convenient to regard a signal as being generated not by a source of e.m.f., but by a current source. Norton's theorem tells us that any linear circuit can be replaced by a current source in parallel with a single impedance, as in Figure 8.11(b). The linear relation satisfied by the Norton equivalent circuit is

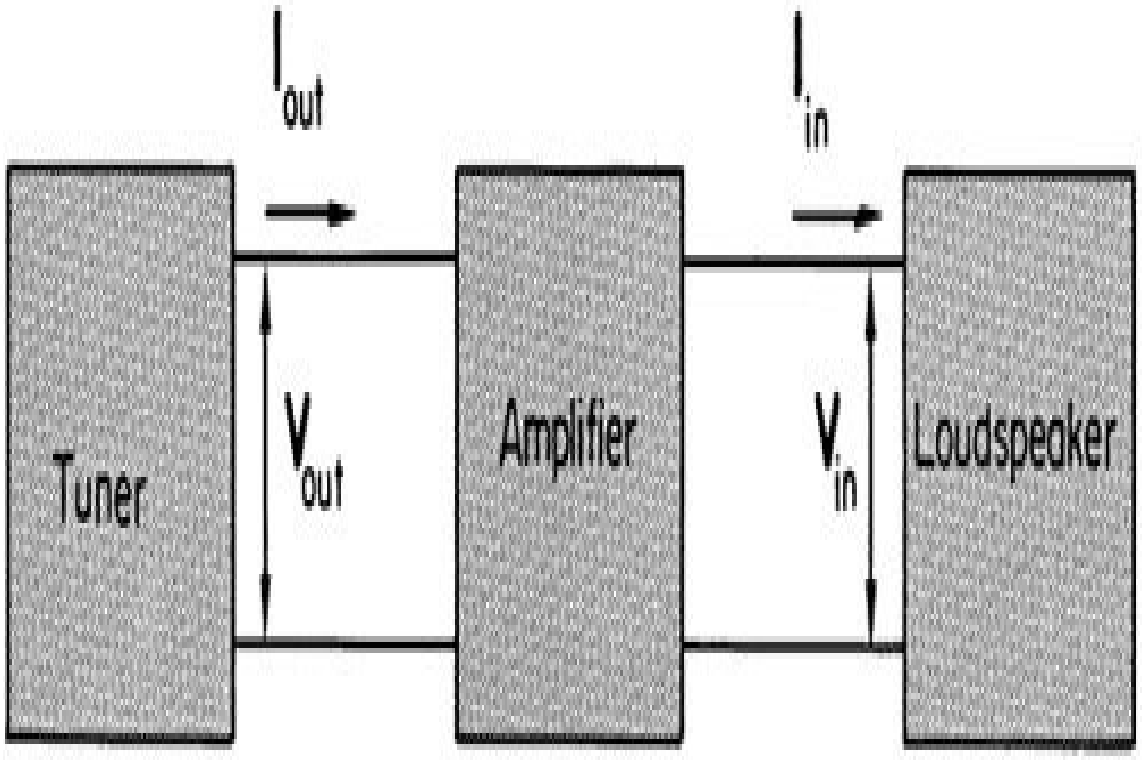


Comparing this equation with equation (8.14), we see that I0 = V0/Zout. The Norton equivalent circuit is particularly useful for circuits containing junction transistors, since these may be regarded as current amplifiers.

If we put Iout = 0 in Equation (8.14), then Vout = V0. When no current is drawn from a circuit, it is said to be an open circuit. The voltage generator appearing in the Thévenin equivalent circuit is thus the open circuit voltage of the real circuit. Similarly, if the terminals on the real circuit are short-circuited, the current is equal to I0, the current source of the Norton equivalent circuit. Since Zout = V0/I0, the output impedance of a circuit may be regarded as the ratio of its open-circuit voltage to its short-circuited current. A word of caution should be added here. Active circuits are rarely linear except within a restricted range, and short-circuiting them is not a recommended practice! However, the equivalent circuits are by definition linear, and it can be helpful in understanding their properties to imagine that they can be open- or short-circuited. But, as for the input impedance discussed in the previous section, it is more rigorous to define the output impedance in terms of signals which are small changes to a steady state, as

(8.15)

Figure 8.12. Replacing preceding and folowing circuits by their output and input impedances.


I_{out}
I_{in}
Tuner
V_{out}
Amplifier
V_{in}
Loudspeaker

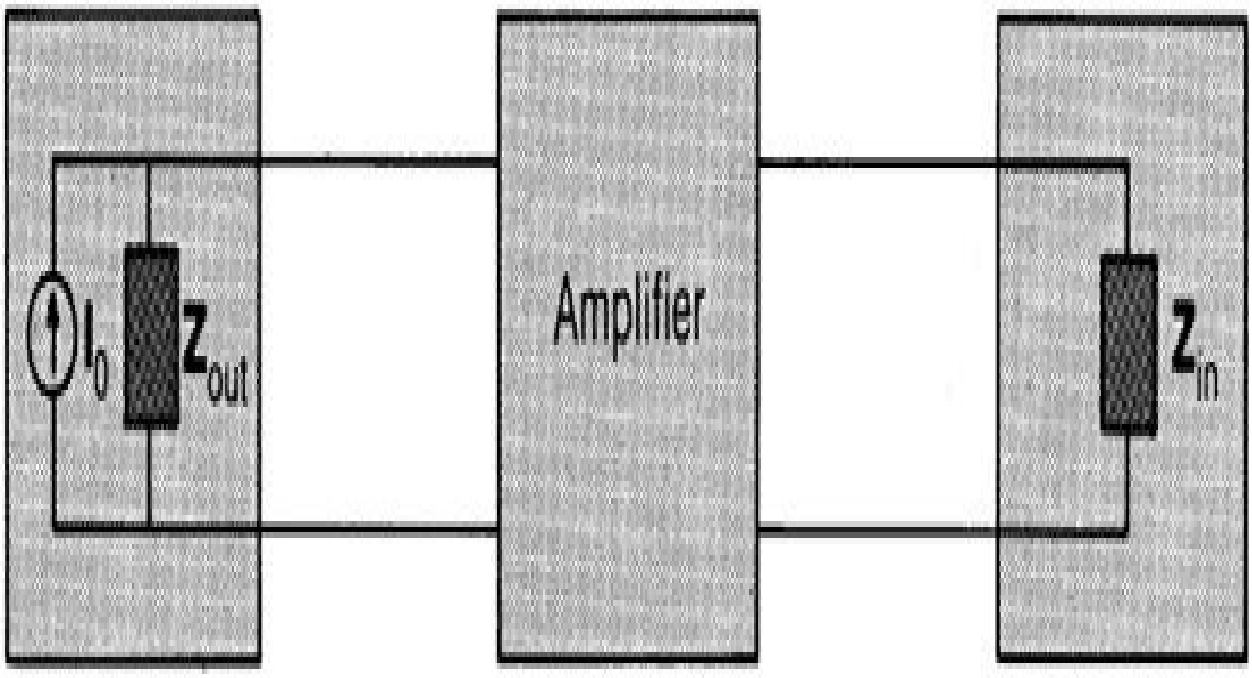

I_0
Z_{out}
Amplifier
Z_{in}

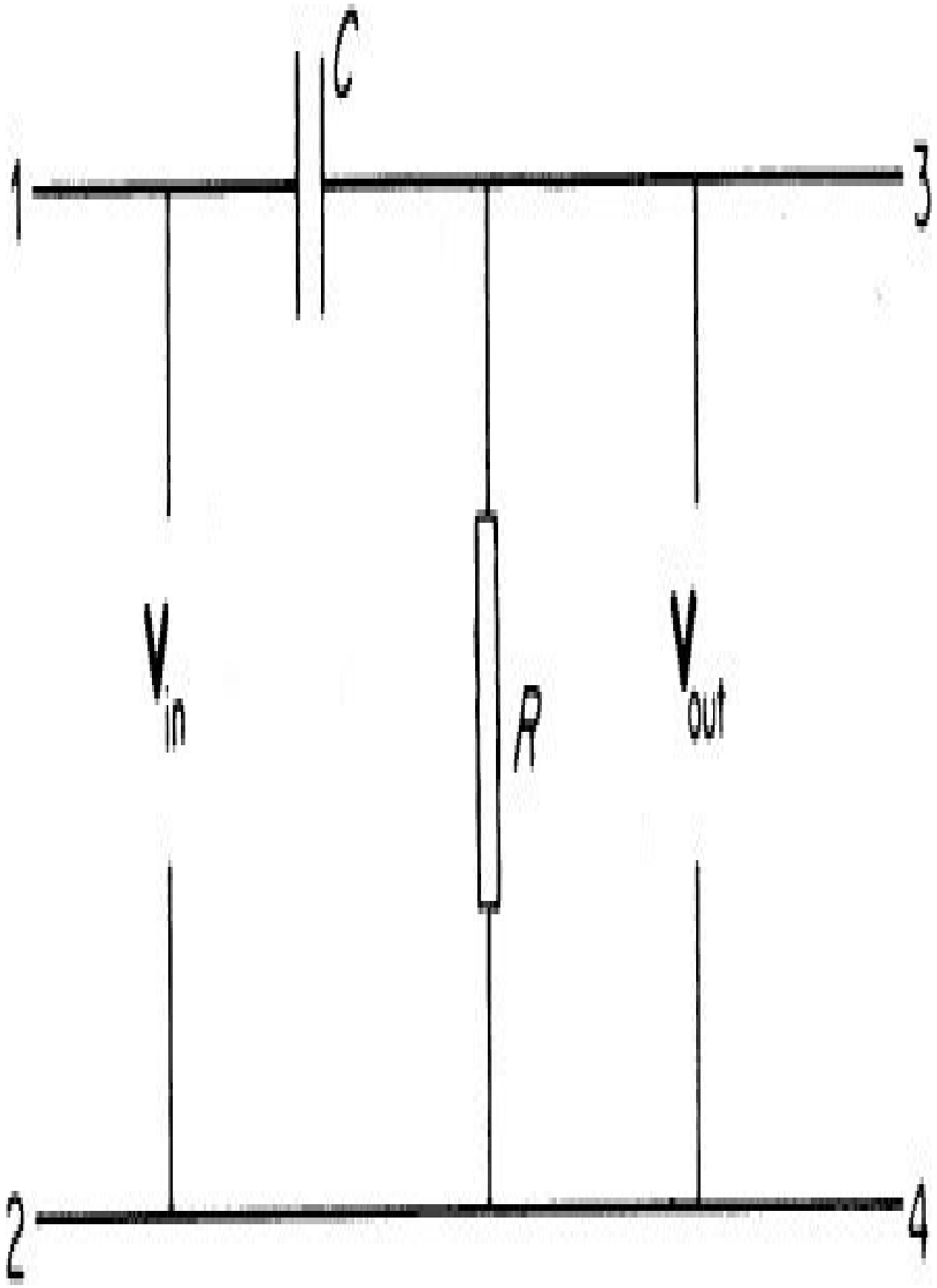
Figure 8.12 shows a schematic diagram of a radio set, divided up into three sections. Each section is a ‘black box’ with input and/or output terminals, and each one can be substituted by an equivalent circuit when analysing its effect on another network to which it is connected. Thus in the lower part of the figure, the amplifier is considered on its own, connected to the equivalent output circuit of the tuner and to the equivalent input circuit of the loudspeaker.

The concepts of input and output impedance have now been introduced in a very general way, so that they are applicable to active circuits. We shall not discuss in detail the input and output impedances of practical active circuits; this topic is left to specialized books on electronics. In the remaining sections of this chapter, however, input and output impedances are calculated for some important passive circuits.

8.4 FILTERS

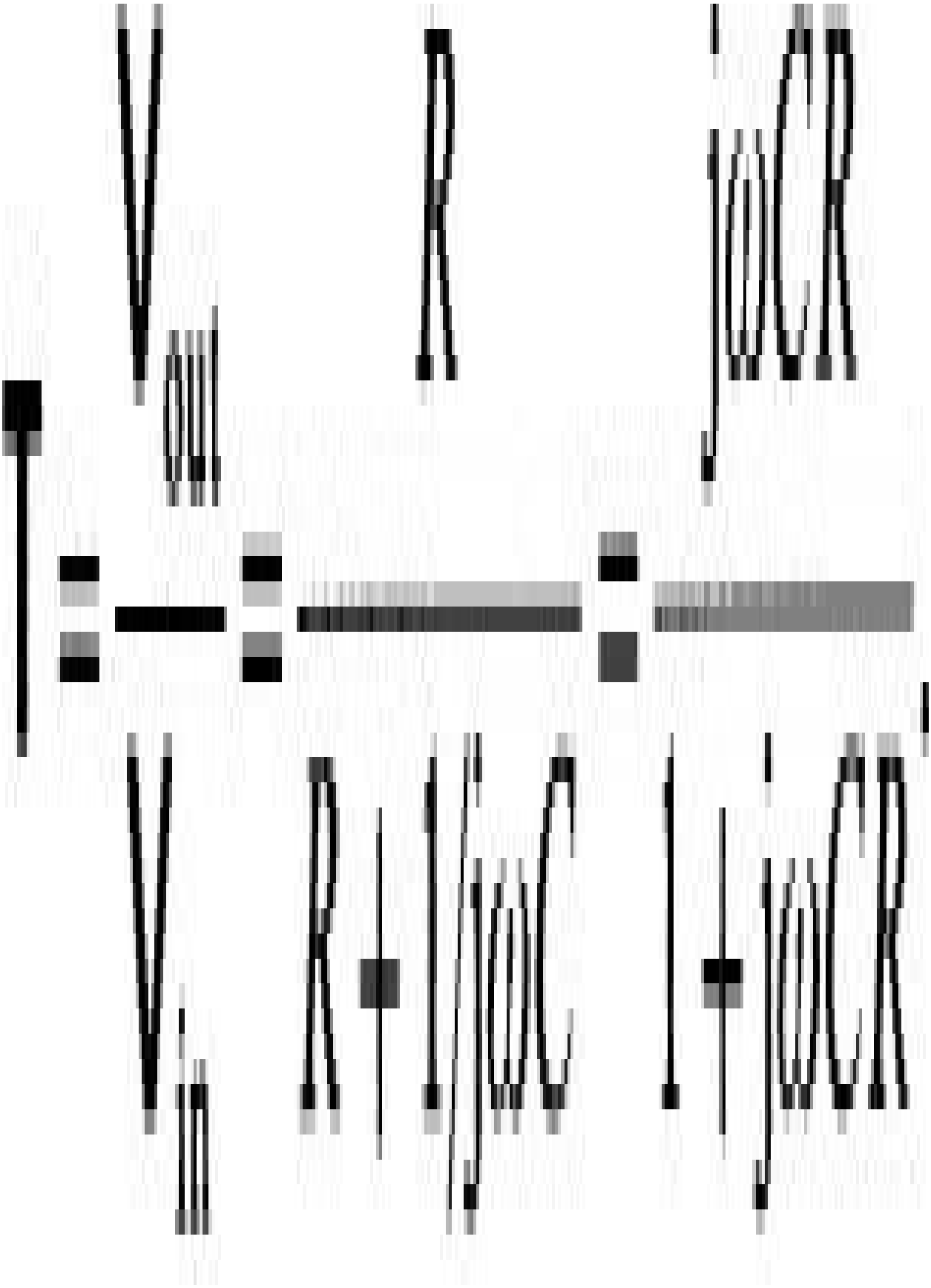
Two sections of an A.C. network are often linked by a capacitor, which transmits alternating voltages while allowing steady voltage differences to be maintained between the sections. Let us consider the behaviour of the network in Figure 8.13 at different frequencies. Across the input terminals 1 and 2 there is a voltage Vin of angular frequency ω. This voltage is modified by the network and a voltage Vout, also of angular frequency ω, appears across the output terminals 3 and 4. The ratio T = Vout/Vin is called the voltage transfer function of the network. The value of T depends on the input impedance of any circuit to which terminals 3 and 4 are connected: we shall assume that this input impedance is so high that a negligible current is drawn from the network, and the transfer function is practically the same as if the network were open circuited. The network then acts as a voltage divider, and since the impedance of the capacitance

Figure 8.13. A simple high-pass filter network.


C
1
3
V_{in}
R
V_{out}
2
4

C in Figure 8.13 is l/jωC, we have

(8.16)



In passing through the network the input voltage is attenuated and also suffers a phase change. Looking at Equation (8.16), we see that at frequencies low enough so that $\omega CR \ll 1$, T is approximately $j\omega CR$. The output voltage is therefore heavily attenuated, and is in advance of the input by a phase angle $\pi/2$. At high frequencies for which $\omega CR \gg 1$, on the other hand, the transfer function is almost unity. The network thus acts as a high-pass filter, transmitting high-frequency signals and attenuating those of low frequency.

Similarly low-pass filters can be made with the simple voltage divider networks shown in Figure 8.14. The open circuit voltage transfer functions for these networks are

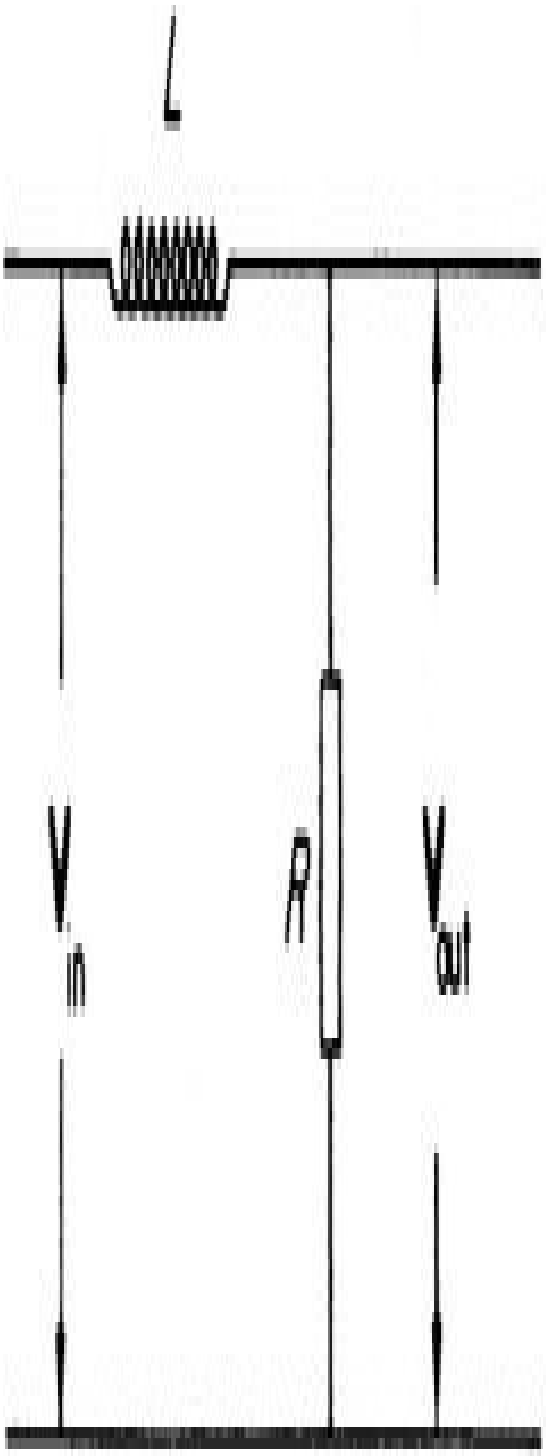
(8.17a)

for the network (a) in Figure 8.14, and

(8.17b)

for the network (b). Both these transfer functions have the same form of dependence on ω. At low frequencies T approaches unity, and the network passes an alternating voltage without modification. At high frequencies, when $\omega L \gg R$ or $\omega CR \gg 1$, as the case may be, the output is attenuated and lags behind the input by the phase angle $\pi/2$. Notice that the input impedances of the two circuits are not the same; (a) is inductive and (b) is capacitative.

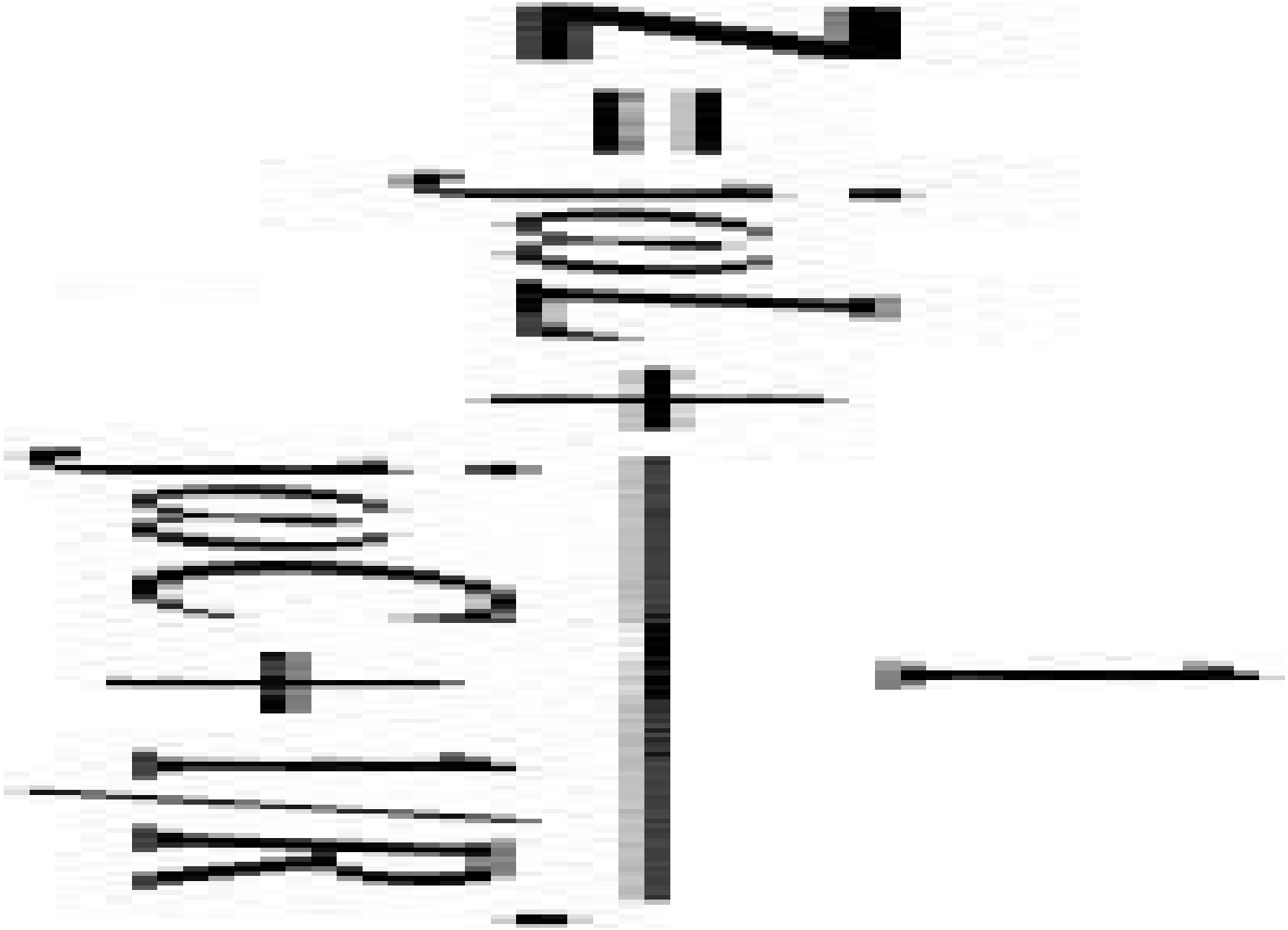
Figure 8.14. Low-pass filters.



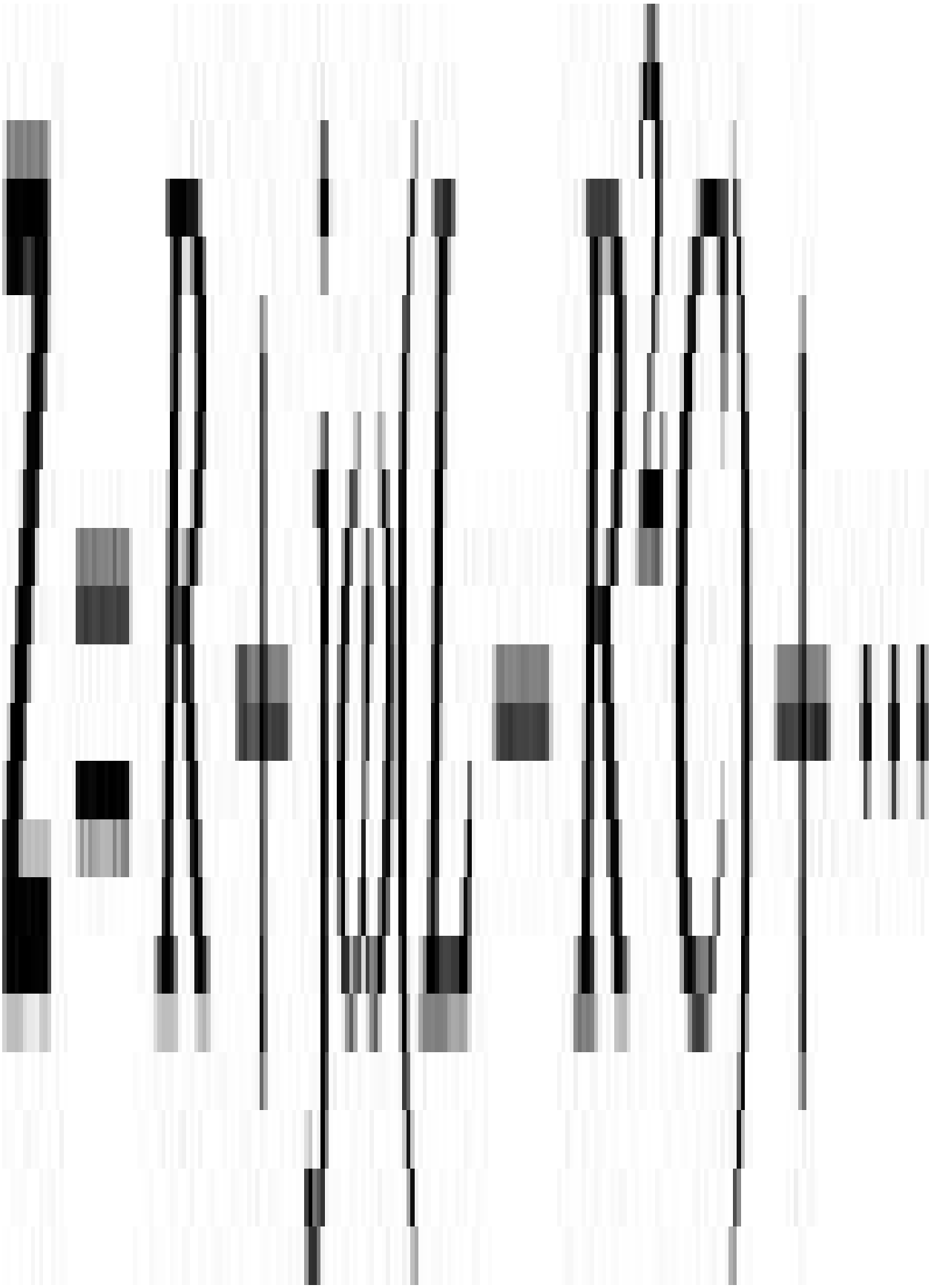
(a) (b)

8.4.1 Ladder networks

Because of their simplicity, voltage dividers are very often used as filters when it is acceptable that the transfer function changes slowly with frequency. But sometimes filters must be designed to meet the more critical requirement that the transfer function should change rapidly from small values to nearly unity over a narrow band of frequencies near some specified cut-off frequency. A sharper cutoff can be achieved by adding more components to the filter network. For example, we can improve the low-pass filter in Figure 8.14(a) by using a capacitor to by-pass the resistor at high frequencies, as shown in Figure 8.15(a). The input impedance of this network is

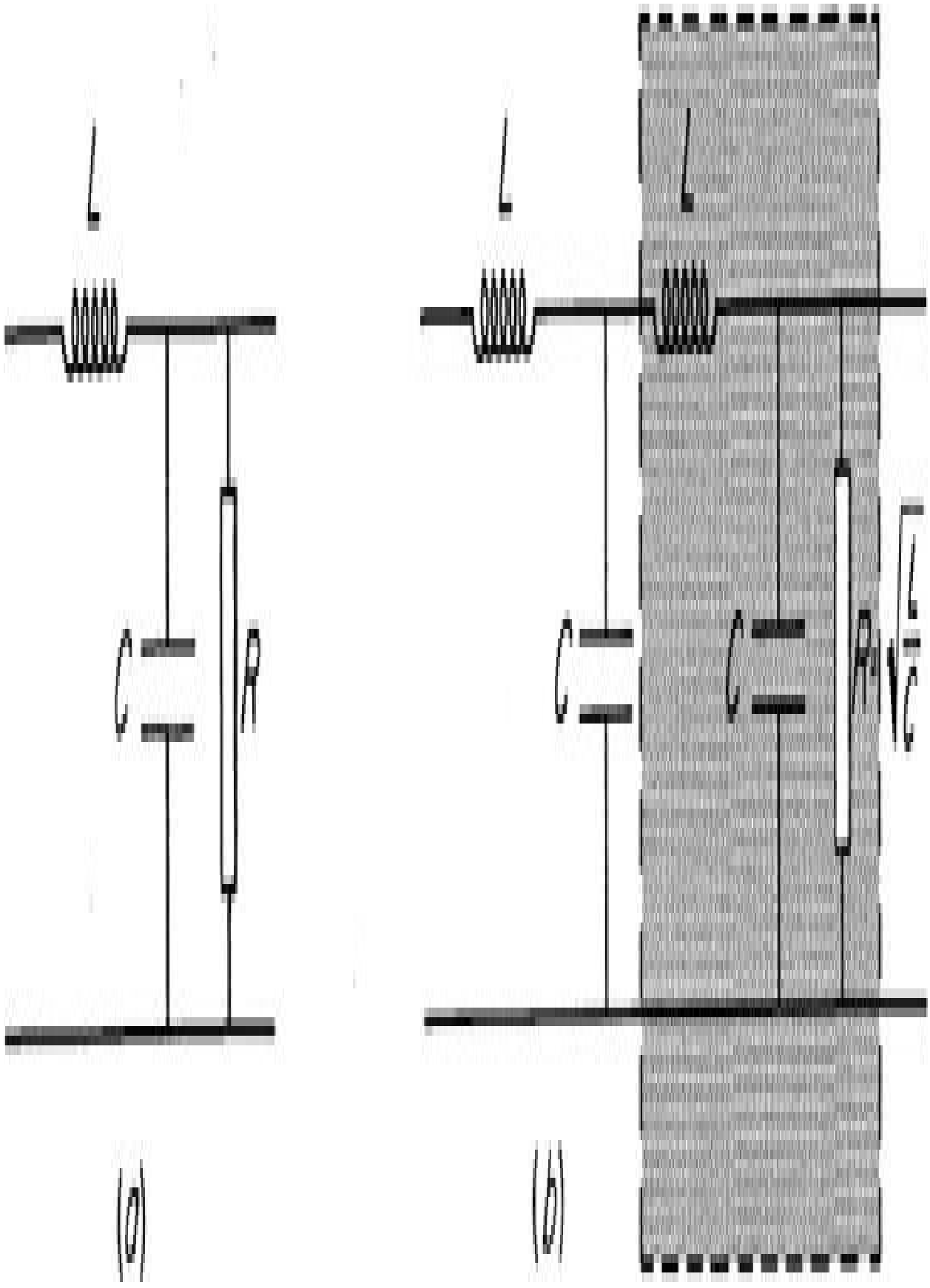


Expanding in powers of ωCR,



If we choose the component values so that $L = R^2C$, i.e. $R = $, then to first order in ω the input impedance is R, the same as if the inductor and capacitor had not been included in the circuit. Now look at the circuit in Figure 8.15(b). The shaded part of the circuit within the dotted lines is just the circuit (a) and at low frequencies is equivalent to a resistance $R = $. Replacing the dotted section by a resistance we are back to circuit (a), and evidently the input impedance of circuit (b) is also at low frequency. Thus we can add more and more LC sections, whilst the low frequency input impedance remains the same. A network of repeating units like this is called a ladder network. When the input impedance of a ladder network at a particular frequency is independent of the number of sections, the network is said to be matched at this frequency, and its input impedance is called the characteristic impedance for this frequency. Thus at low frequencies is the characteristic impedance of the LC ladder.

Figure 8.15. A low-pass filter made with a ladder network of similar sections.

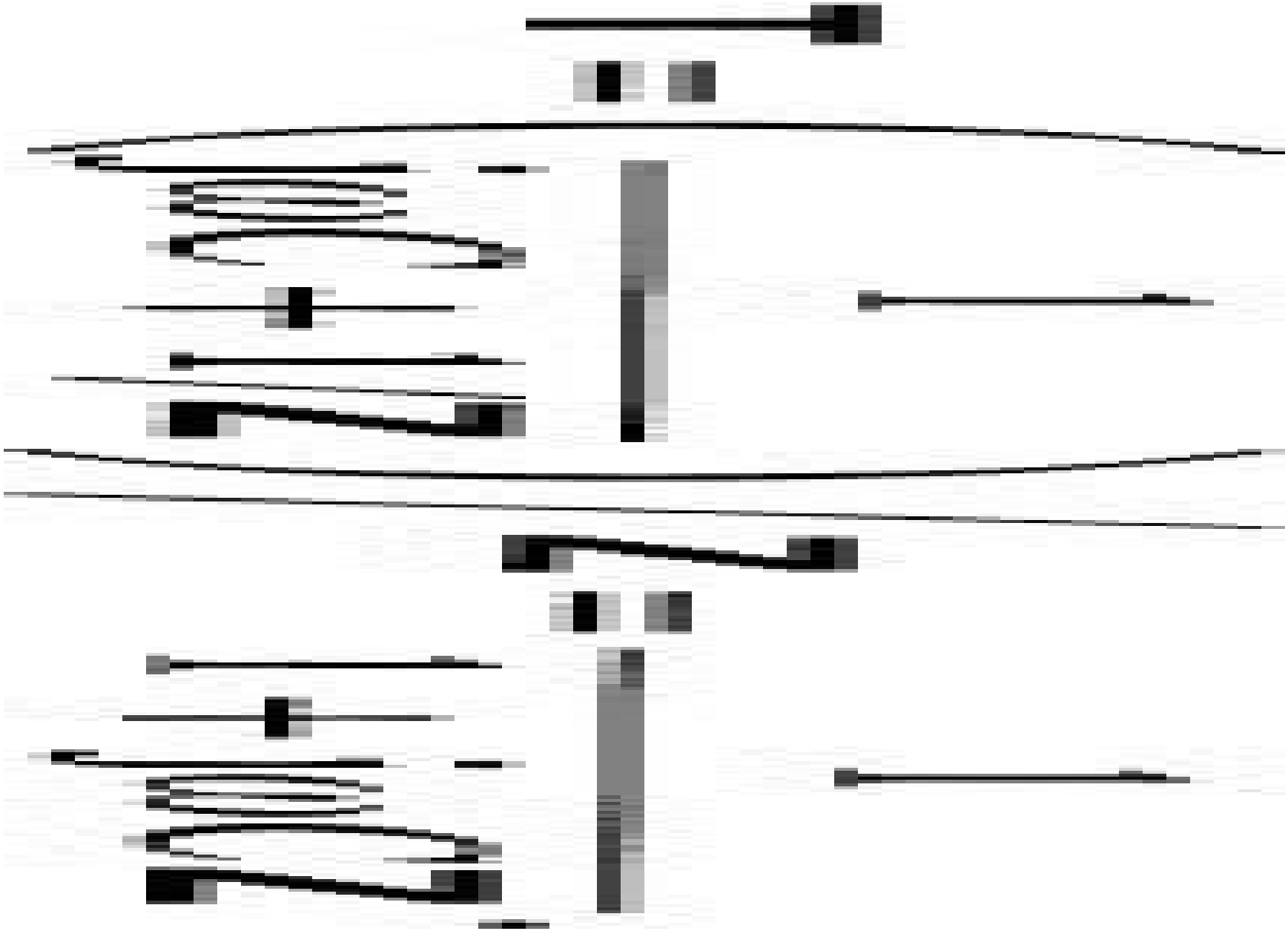

L
C
R
L
C
RN
C

To find out how efficient the ladder is as a low-pass filter, we must work out its transfer function at different frequencies. Imagine for simplicity that the ladder is matched at all frequencies, i.e. it is terminated by an impedance Z which is equal to the characteristic impedance of the ladder at all frequencies. The impedance Z is approximately equal to at low frequencies, but it includes terms in ω^2 and higher powers of ω which become important at high frequencies. The input impedance of the network in Figure 8.16, formed by adding a single section to an existing ladder, must also be Z:

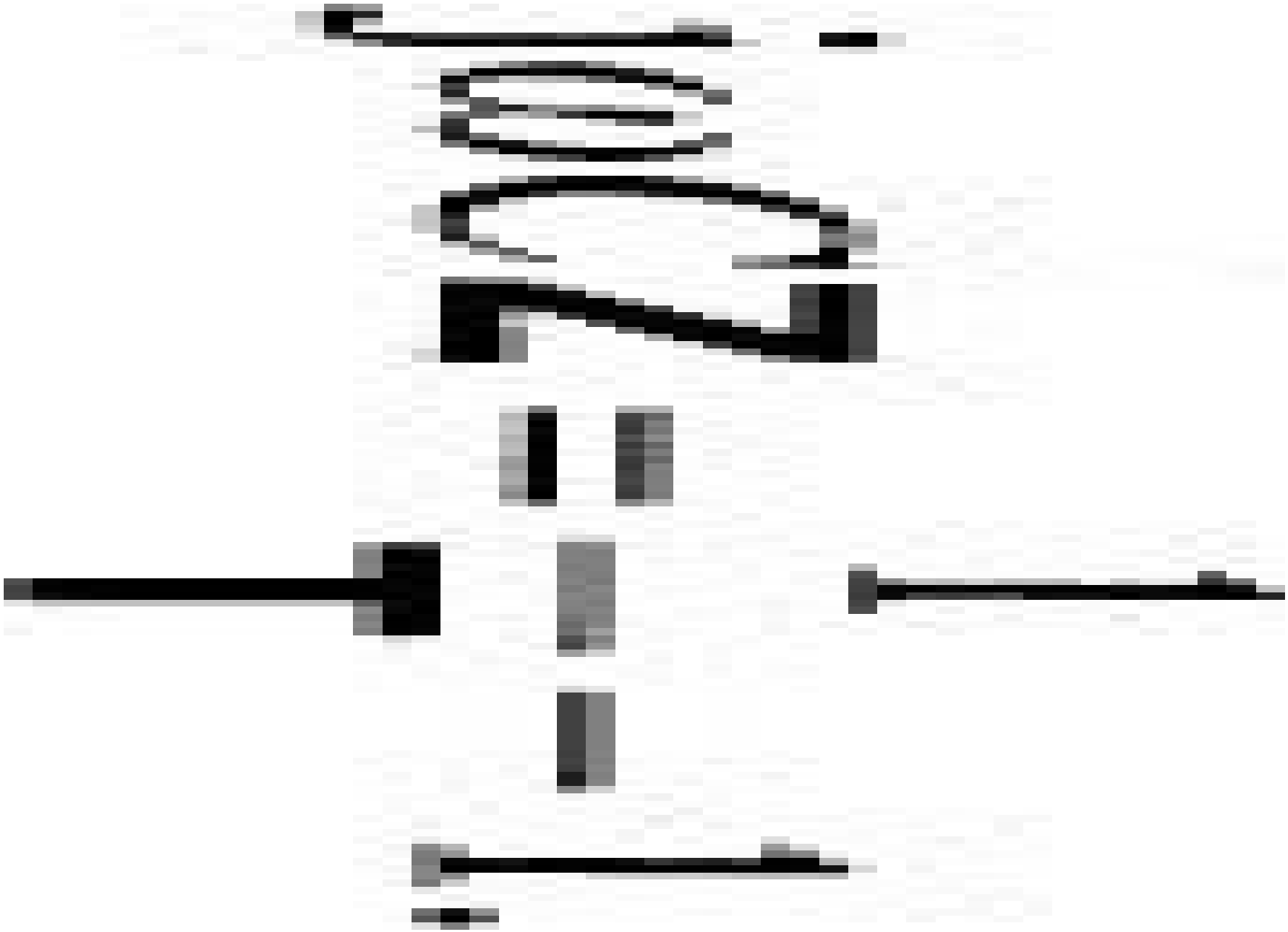
hence

(8.18)

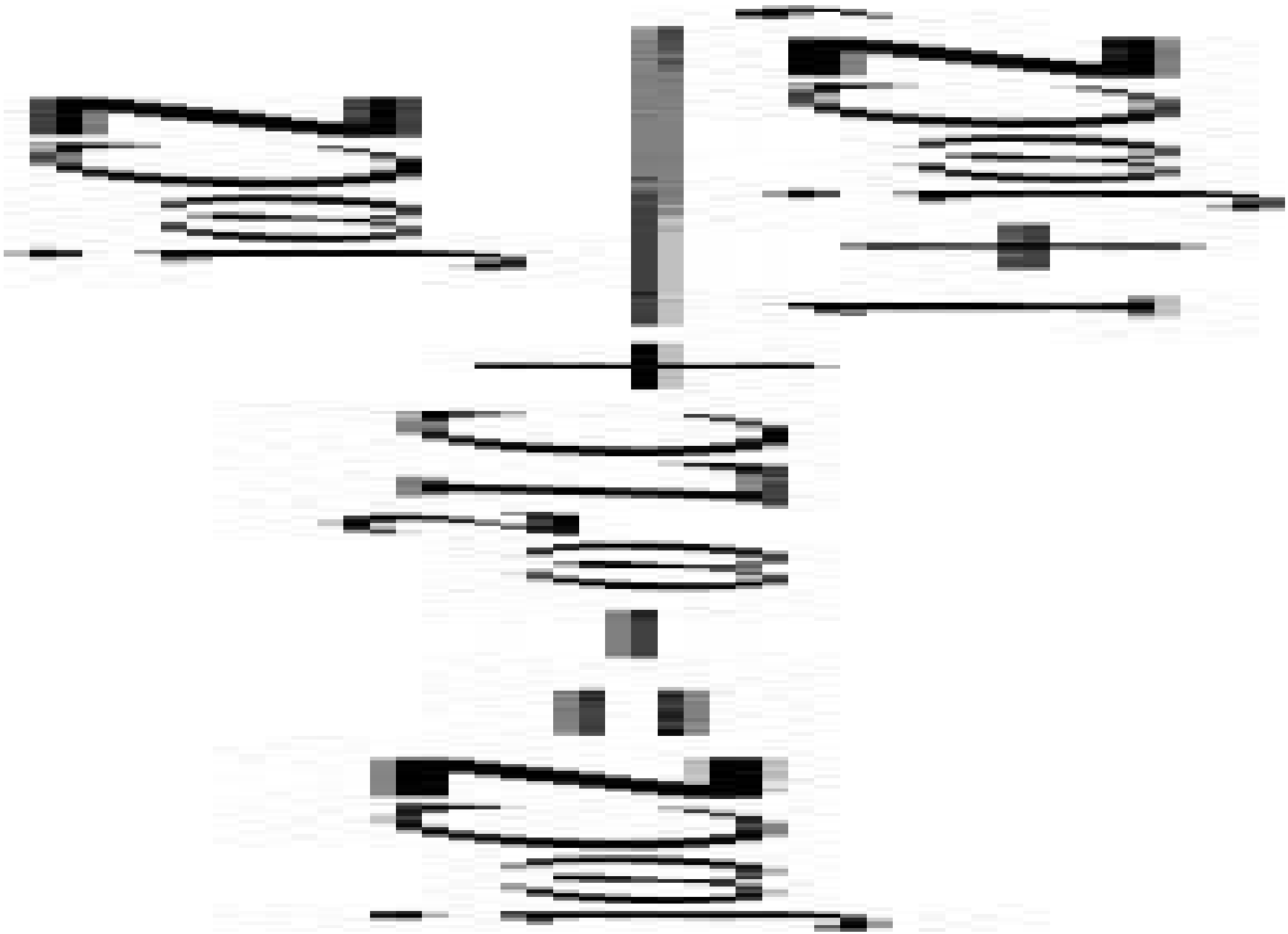
The voltage transfer function of the new section is



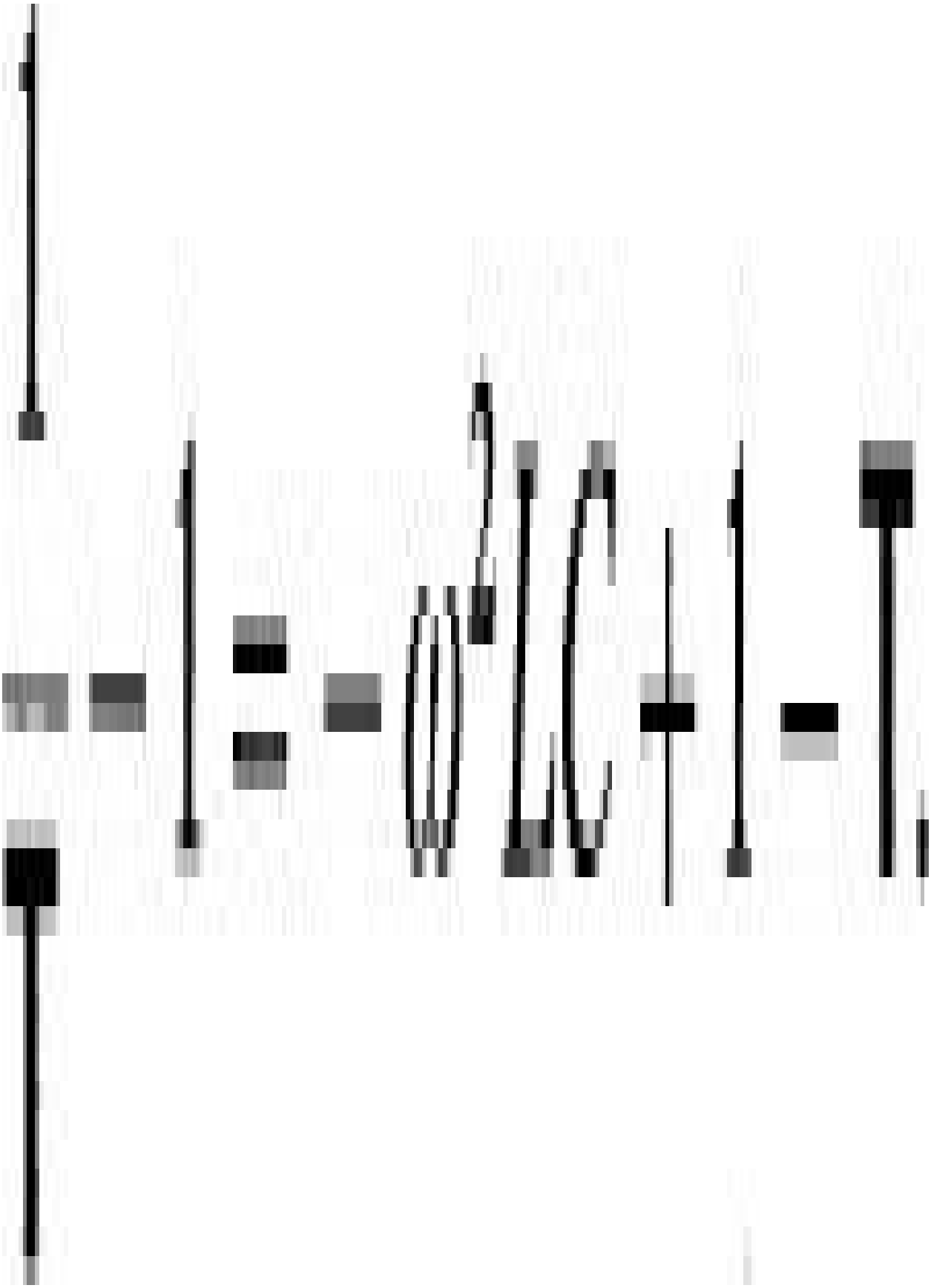
Hence



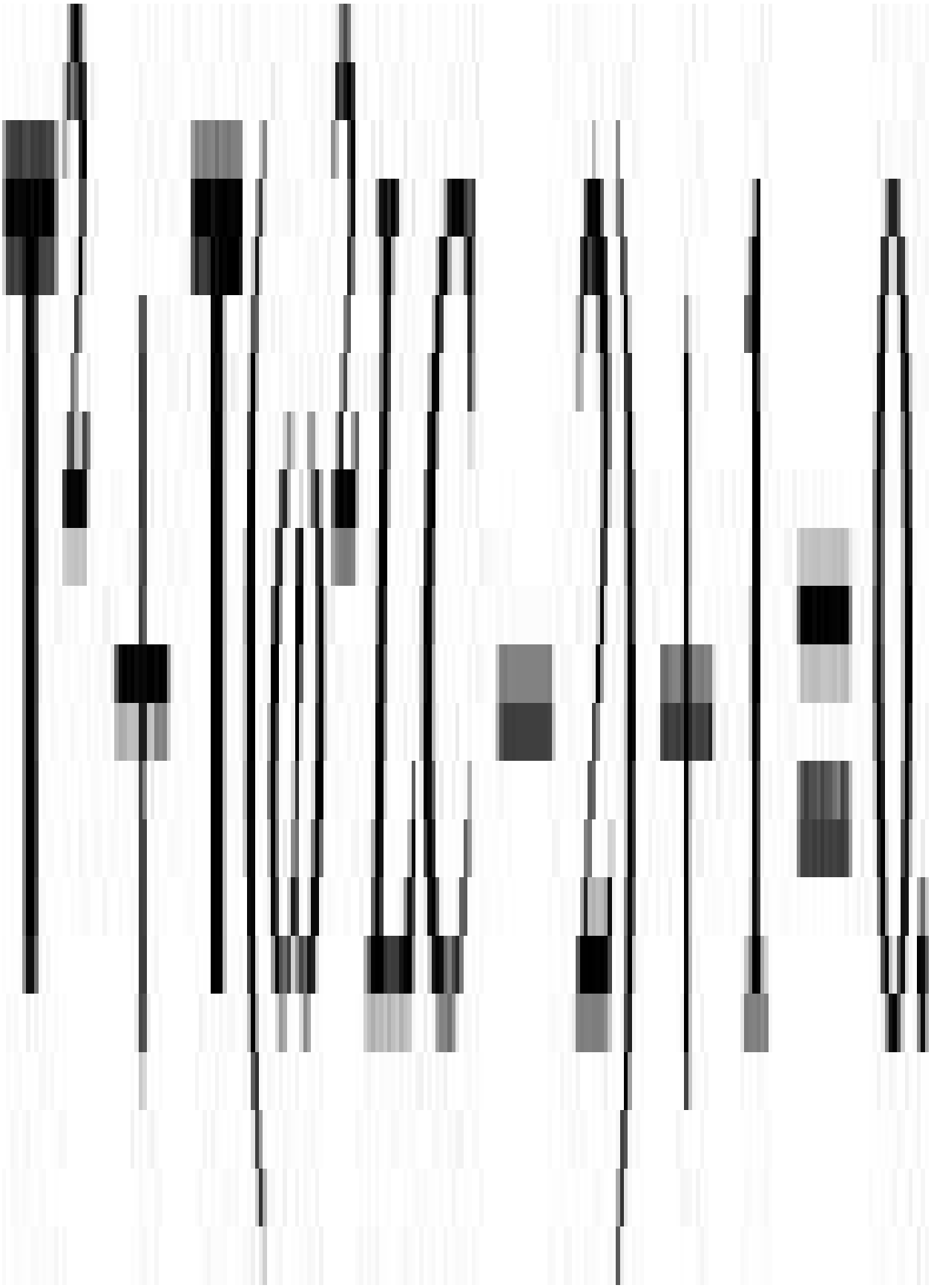
From 8.18,



or



Rearranging,



The solution is

(8.19)

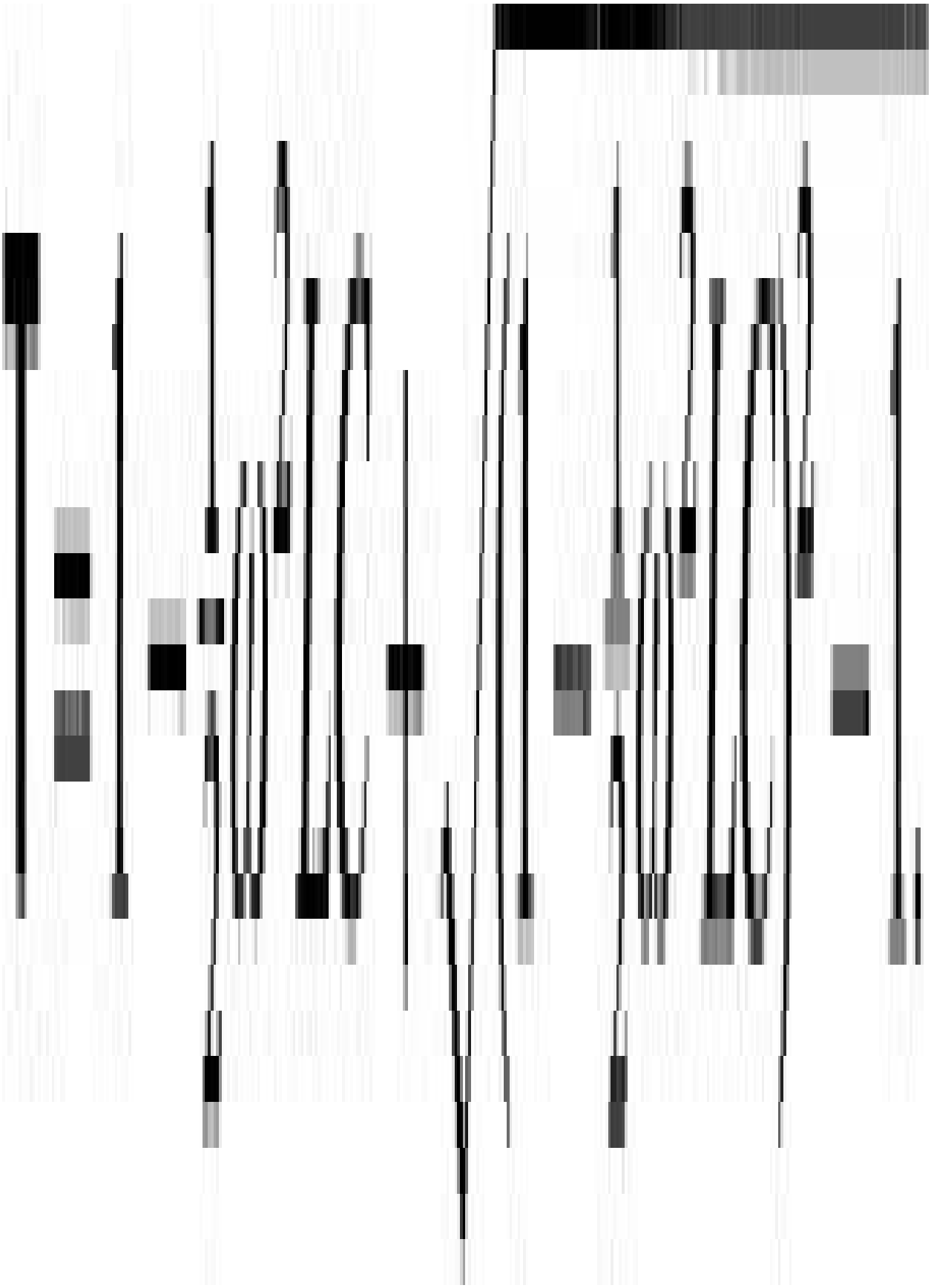
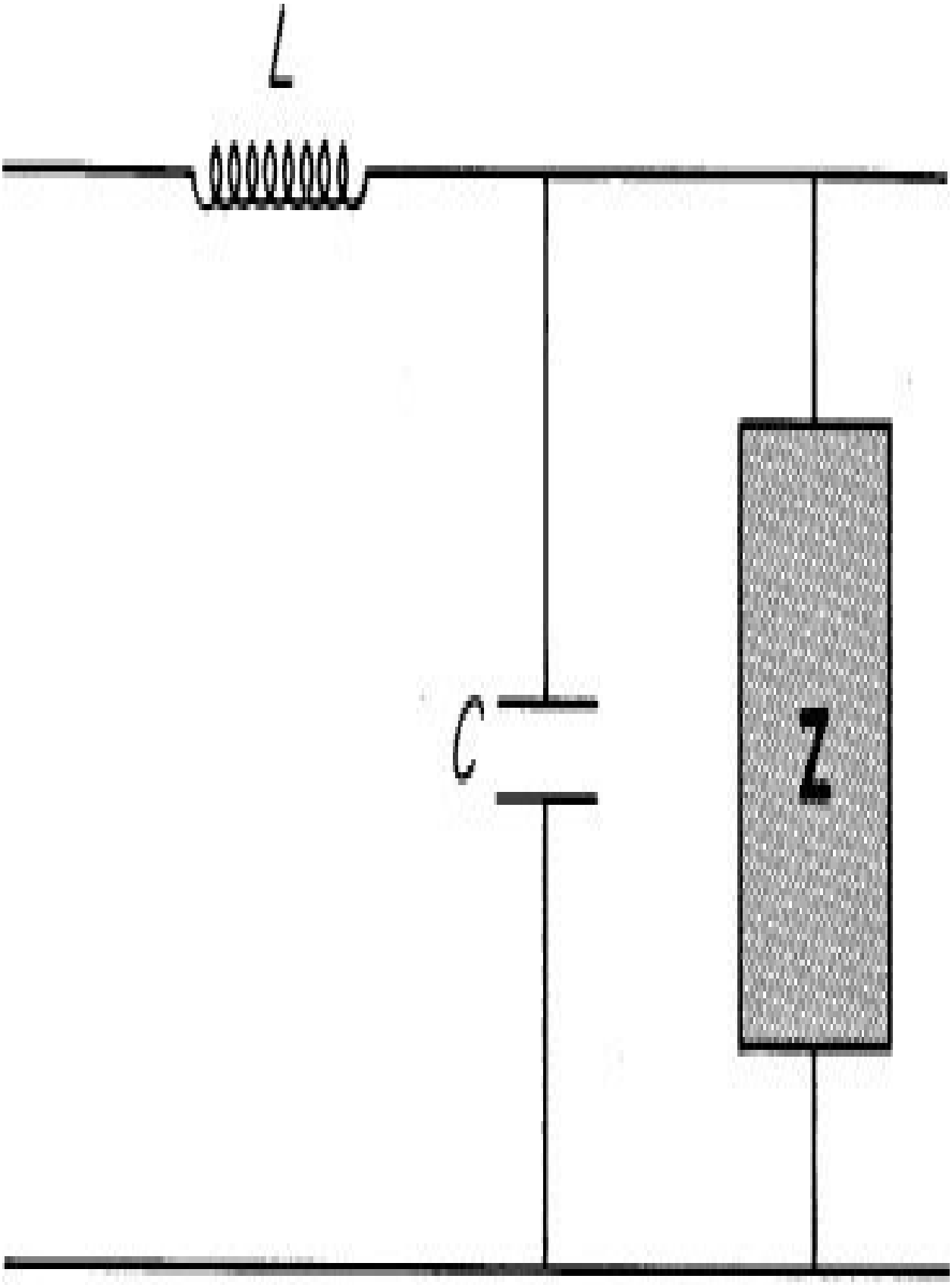
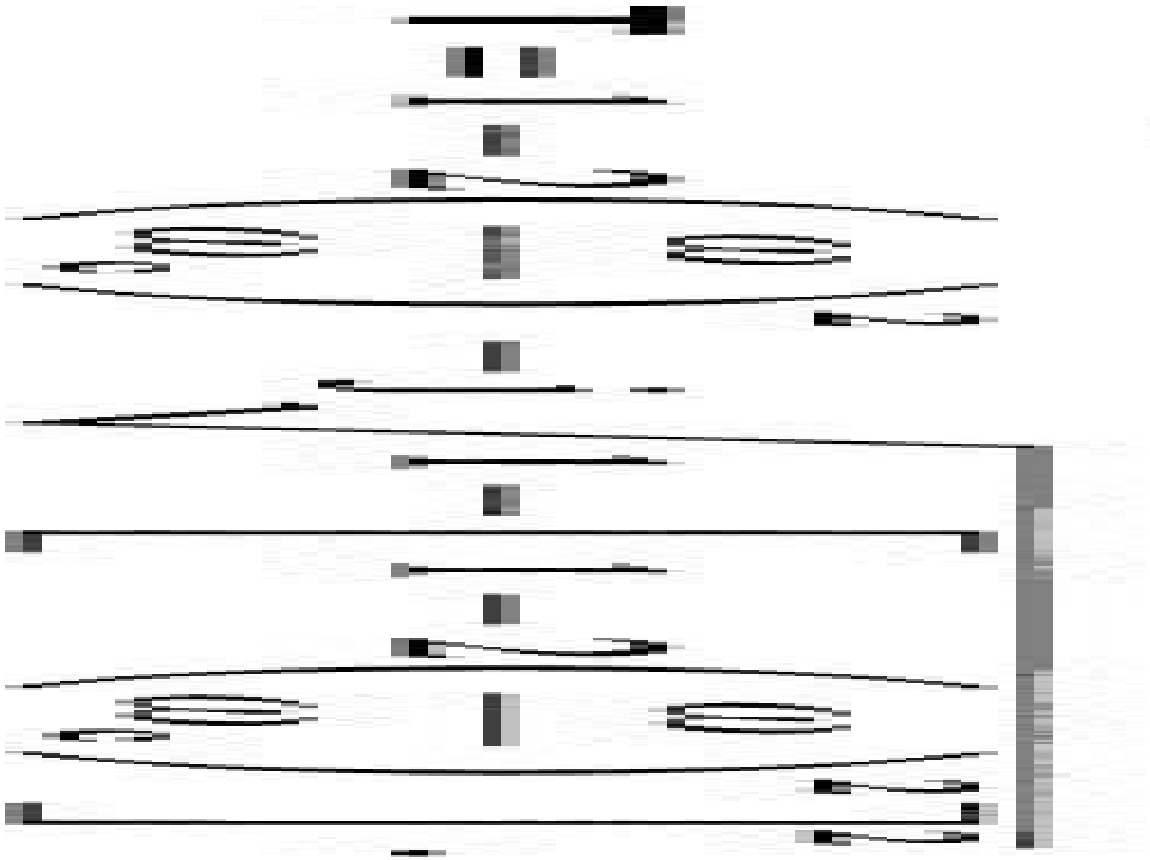


Figure 8.16. An additional section added to a low-pass filter with input impedance Z.


L
C
Z

(We have ignored the solution with a negative sign in front of the square root: this solution represents a reversal of the role of input and output, when a signal travels backwards through the network.)

The behaviour of the network depends on the sign of the quantity under the square root in Equation (8.19). The sign changes at a cut-off angular frequency $\omega_c = 2/$. Let us consider angular frequencies above and below this cut off.



or

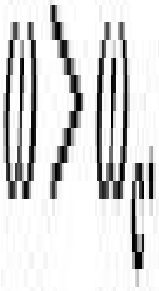
(8.20)

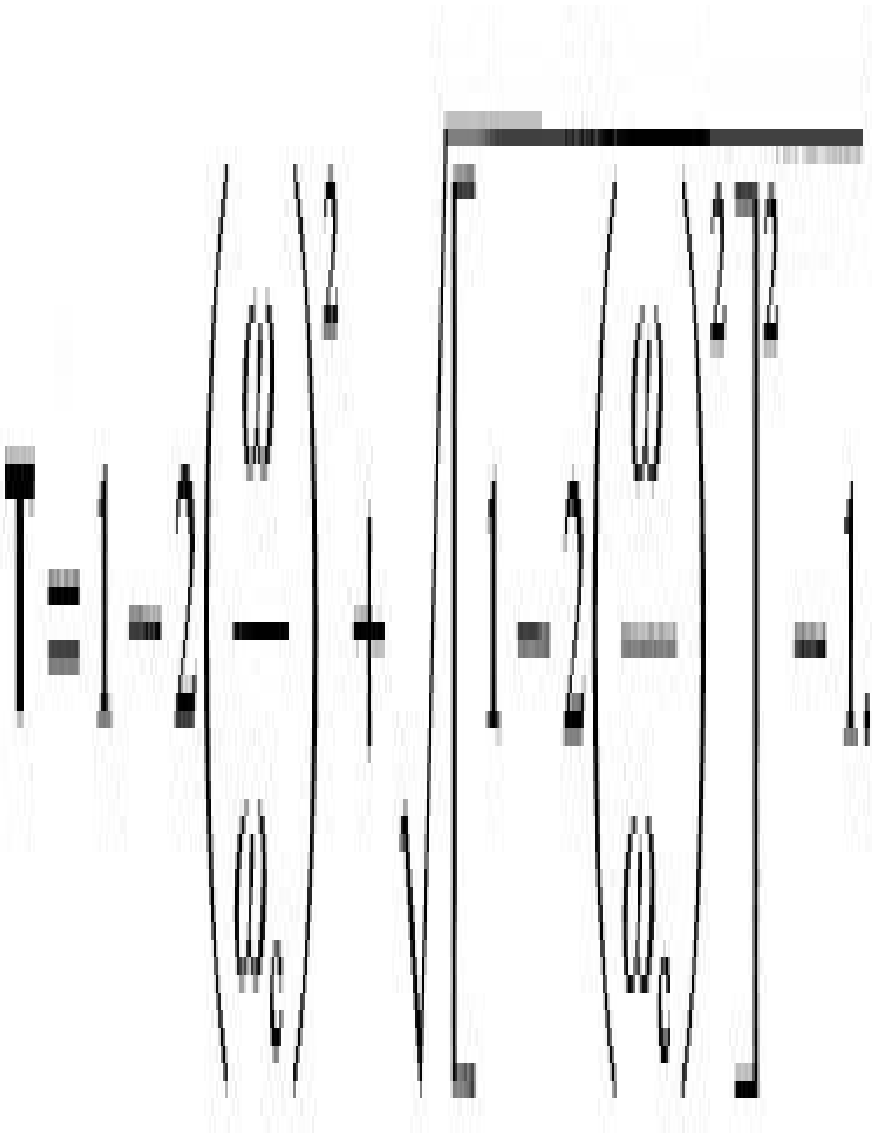
where

(8.21)

The alternating voltage is transmitted without attenuation but in passing through each section of the network its phase is changed by the angle θ.

(8.22)





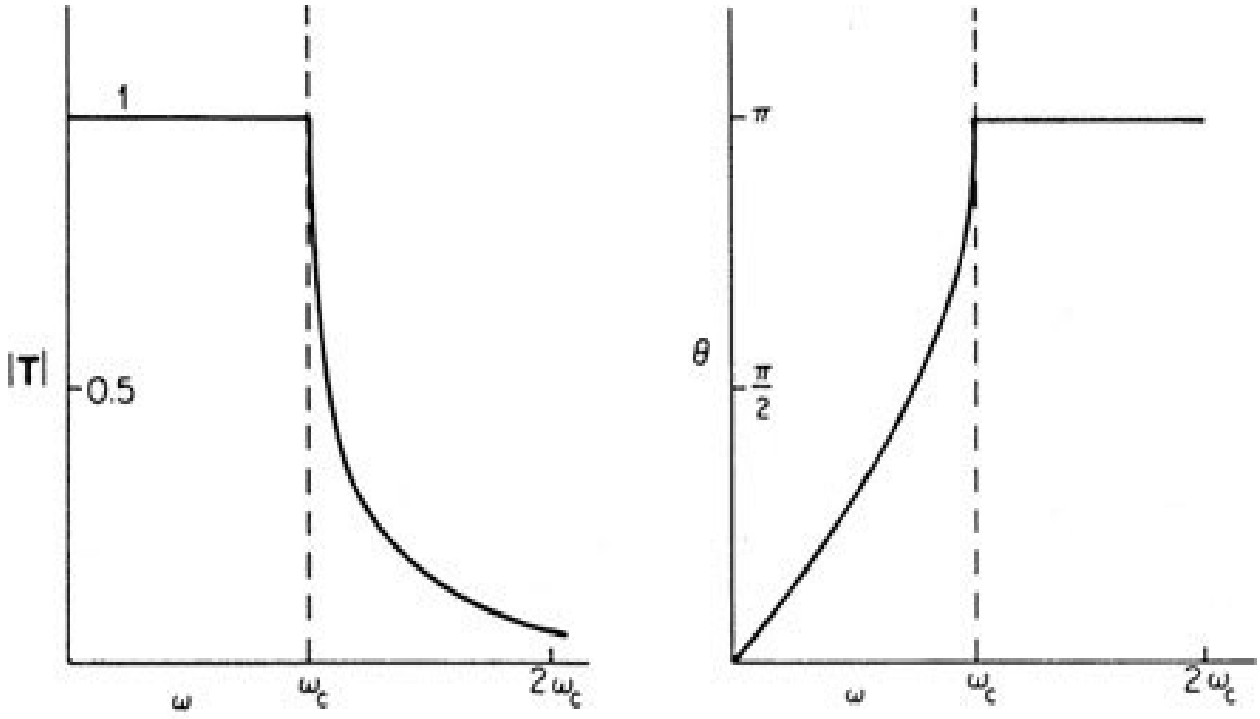
The transfer function is real and it lies between (− 1) and 0. The voltage signal is therefore attenuated and reversed in phase by each section.

The modulus and phase of T are sketched in Figure 8.17. The angular frequency range up to the cut-off at ωc is called the pass-band of the network, and the range above ωc is the stop-band. In practice it is not possible to construy an ideal filter with no attenuation at all in the pass-band, because real coils always have some resistance, unlike the ideal inductances we have used in the equivalent circuits. Further, it is not possible to make a component with impedance equal to the characteristic impedance of the ladder network at all frequencies; such a component is required to terminate an ideal filter. However, there is a reasonably sharp cut-off even for the single-section filter in Figure 8.15(a) if it is terminated by a resistance equal to its low-frequency characteristic impedance : the modulus and phase of the transfer function of such a filter are compared with the ideal in Figure 8.17.

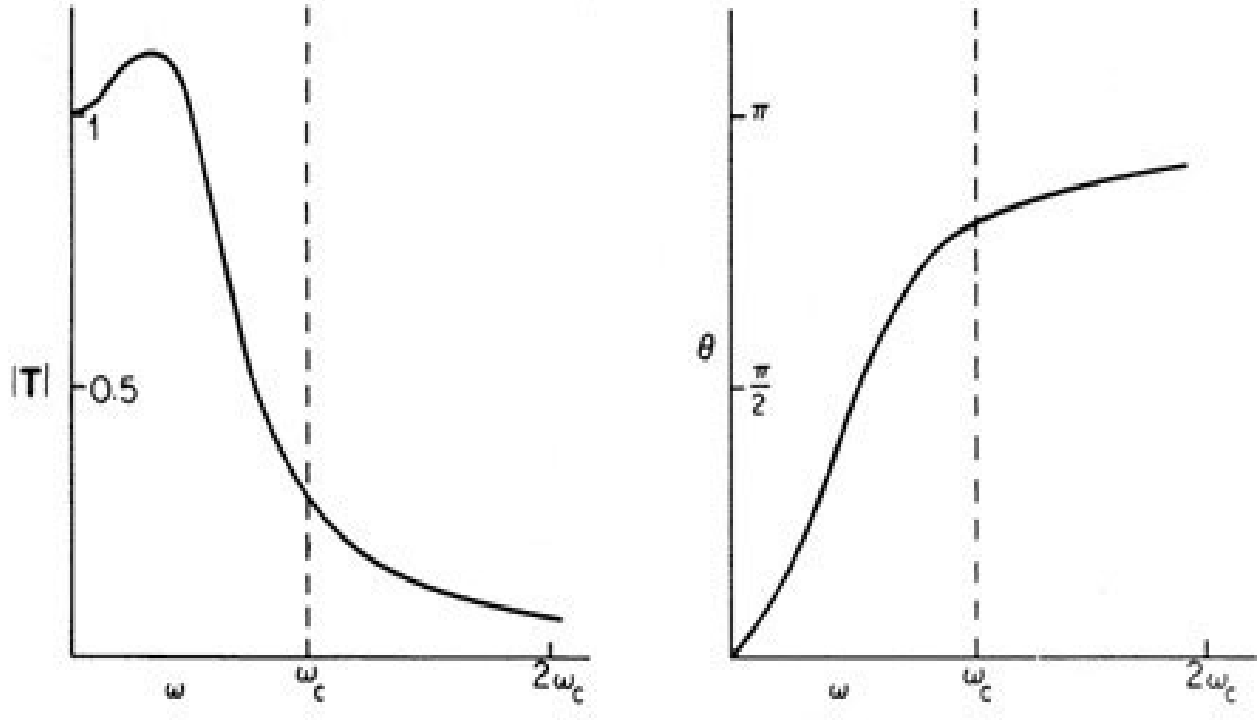
8.4.2 Higher order filters and delay lines

Low-pass filters made up of several sections of a ladder network have much greater attenuation above the cut-off frequency than does the single section filter. As an example, let us design an LC filter to reject unwanted signals at the input of an amplifier. Suppose that the amplifier has an input impedance of 100Ω and that the filter is required to pass a signal at a frequency of 1 MHz, but not to allow more than 1% transmission of the 2 MHz harmonic. We choose the cut-off frequency to be 1.5 MHz, so that the 1 MHz signal in the pass-band and the 2 MHz harmonic in the stop-band are both well away from the cut-off. The transfer function of a three-section filter which is matched at low frequencies and has its cut-off at 1.5 MHz is sketched in Figure 8.18. The transmission is 80% at 1 MHz and 0.6% at 2 MHz, so that it meets the requirements of the problem*. The component values of the filter are fixed by the input impedance of the amplifier (100Ω) and the cut-off frequency (1.5 MHz), since we have

Figure 8.17. The magnitude and phase of the voltage transfer function for low-pass filters.

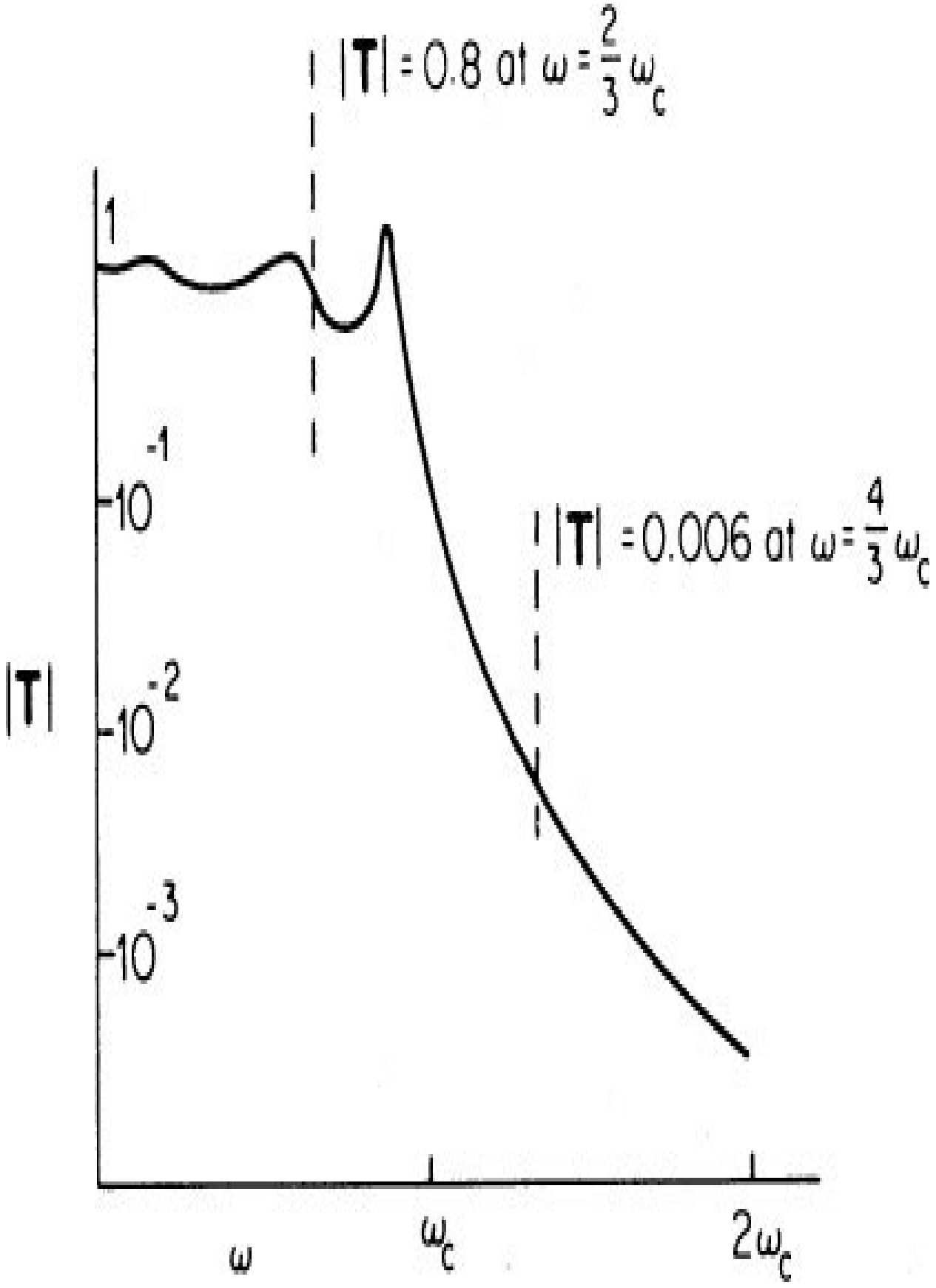


(*a*) A single section of an ideal ladder matched at all frequencies.

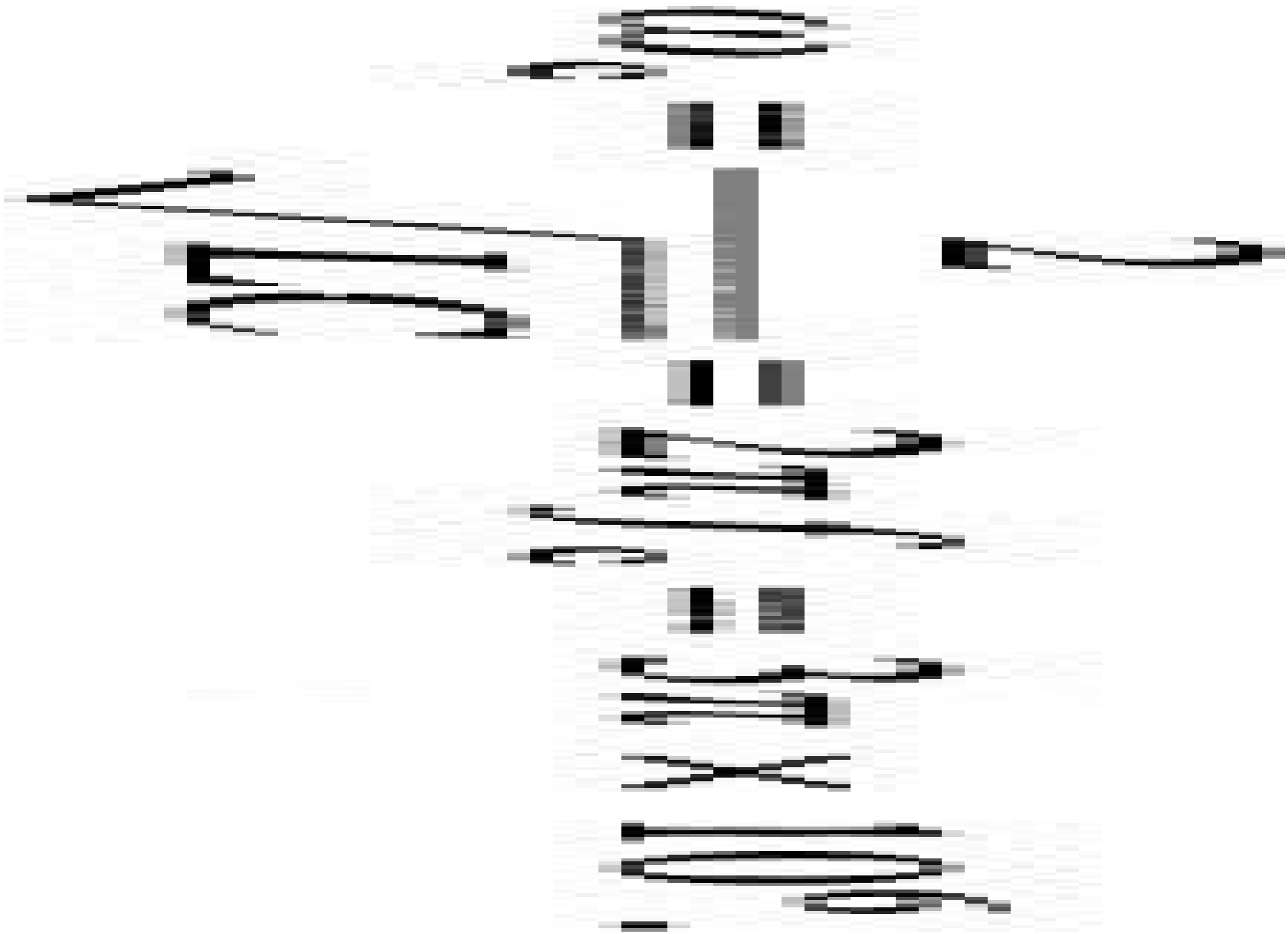


(*b*) A single section terminated by a resistance $\sqrt{\frac{L}{C}}$.

Figure 8.18. The transfer function of a three-section filter matched at low frequencies.


|T| = 0.8 at ω = 2/3 ω_c
|T| = 0.006 at ω = 4/3 ω_c
1
10^-1
10^-2
10^-3
|T|
ω
ω_c
2ω_c

and



The solution of these two equations is L 22 μH, C 2200 pF.

High-pass and band-pass filters

As well as being used as low-pass filters and delay lines, ladder networks can be made up into high-pass or band-pass filters. (A band-pass filter is one which passes signals in a band between two cut-off frequencies, and attenuates signals outside the band.) The analysis of these filters proceeds along the same lines as for the low-pass filter, and we shall not present it here. Networks which act as high-pass and band-pass filters are shown in Figures 8.19 and 8.20.

Figure 8.19. A high-pass filter with characteristic impedance R = at high frequency.

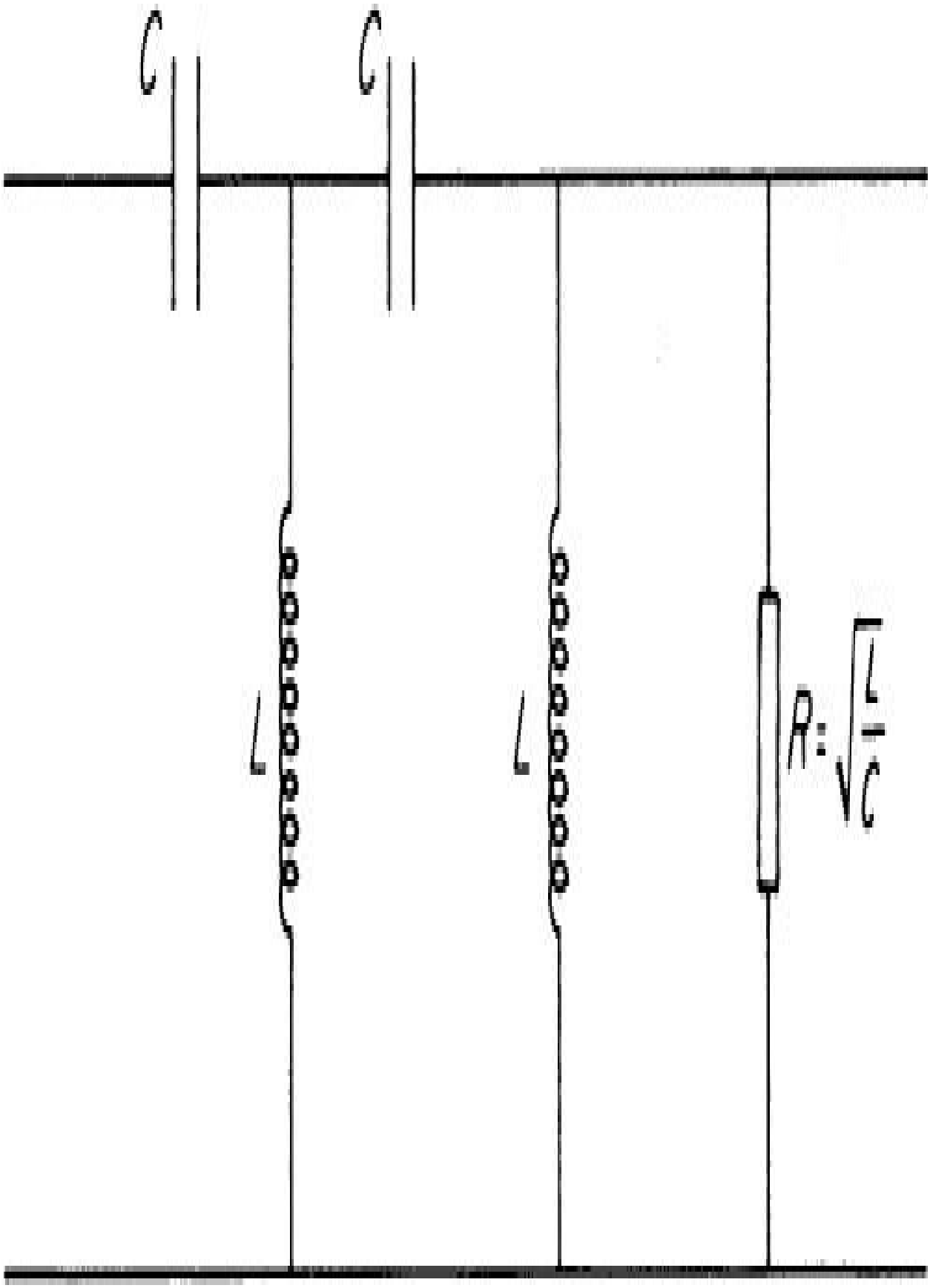
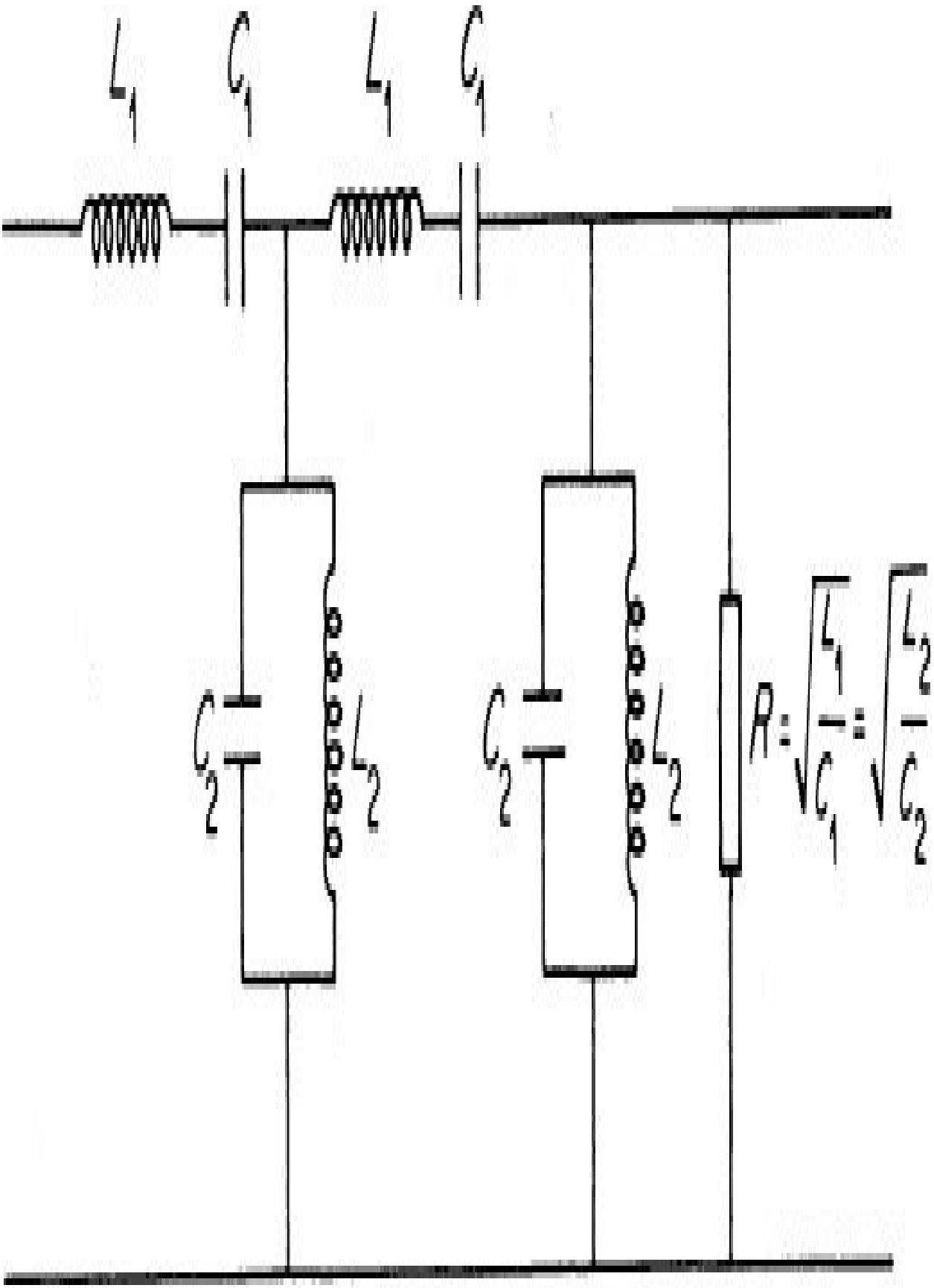
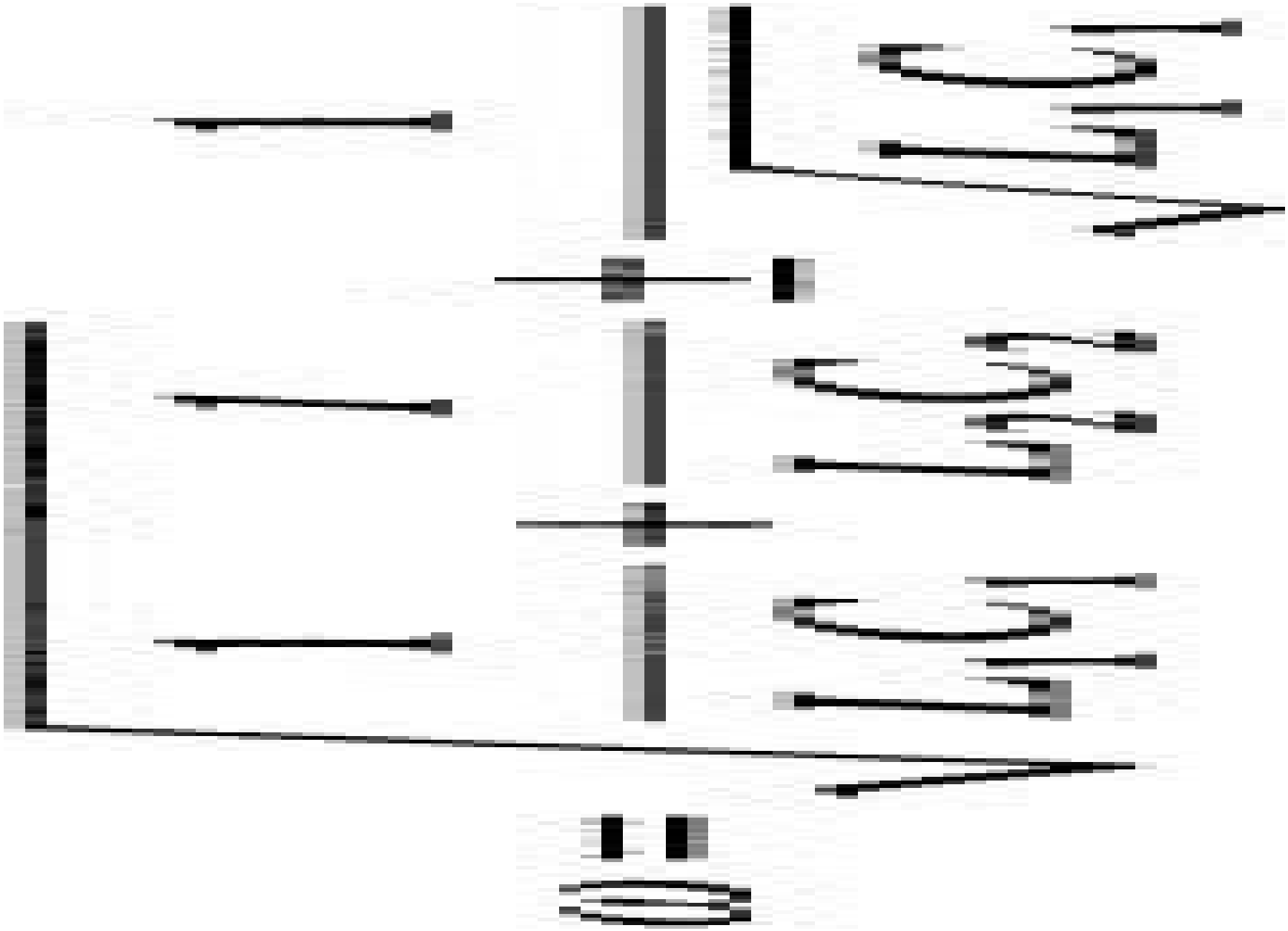

C
C
L
L
R = √(L/C)

Figure 8.20. A band-pass filter. The cut-off frequencies are at

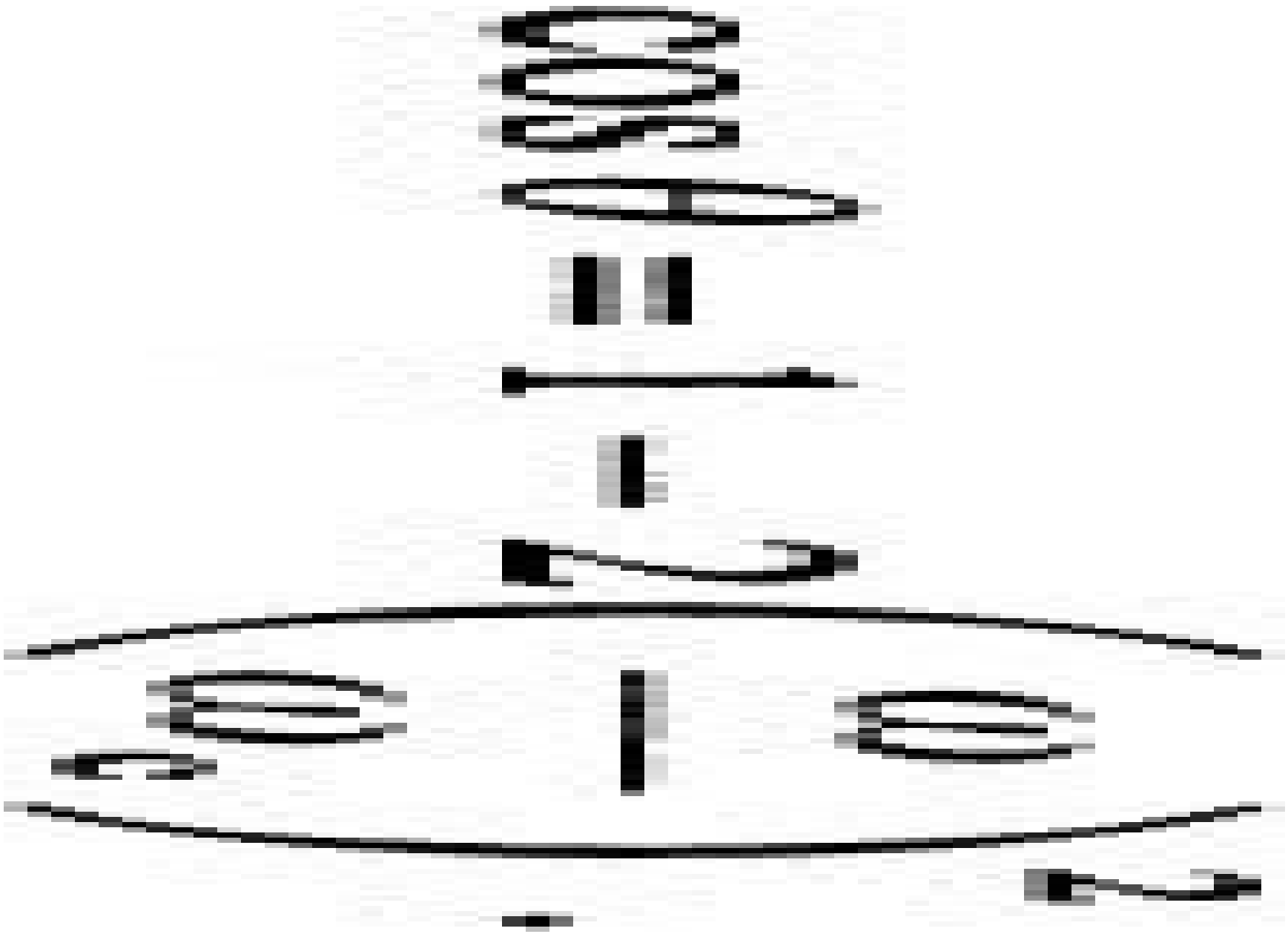

L_1
C_1
L_1
C_1
C_2
L_2
C_2
L_2
$R = \sqrt{\frac{L_1}{C_1}} = \sqrt{\frac{L_2}{C_2}}$



Making the expansions>and the characteristic impedance is R near the centre of the pass-band.

Delay lines

A signal in the pass-band of an ideal matched low-pass filter is not attenuated, but it is changed in phase. The phase lag θ introduced by a single section may be found from Equation (8.21)



Making the expansion $\cos\theta = 1 - \theta^2 + \theta^4 + \ldots$ we see that provided θ^4 is small, θ $2\omega/\omega c$. Each section introduces the same phase shift θ, and the alternating voltage at the input is progressively delayed in phase as it passes along the network. Associated with the phase shift is a time delay; a phase lag of 2π radians corresponds to a time delay equal to the period $2\pi/\omega$ of the alternating voltage. Because the phase shift introduced by each section of the network is proportional to angular frequency for angular frequencies well below cut-off, the

time-delay

(8.23)

is constant. Any signal made up of a superposition of alternating voltages at low frequencies is delayed by the same time in passing through the filter network. Since the low frequencies are not attenuated, the network acts as a delay line, delaying the signal without distorting it.

In practice this kind of delay line is only satisfactory if it does not contain very many sections. The delay of each section is $2/\omega c = 1/\pi fc$, and the maximum delay which can be introduced by a line made up of a few sections is of the order of the period of its cut-off frequency. Looking at the transfer function for the three-section line in Figure 8.18, one sees that the transmitted amplitude begins to change rapidly with frequency at about $\omega/\omega c =$. Above this frequency the time delay also ceases to be constant. For this filter $l/\pi fc = 1/(1.5\pi)\mu s$, and in passing through all three sections, low frequencies are delayed by $2/\pi = 0.64\ \mu s$. The delay at $fc = 0.75$ MHz is 0.67 μs, nearly the same as the low-frequency value, but the delay at 1 MHz is 0.81 μs. Non-sinusoidal signals including frequency components as high as 1 MHz will suffer considerable distortion in passing through the network.

8.5 TRANSFORMERS

A transformer consists of two or more coils wound on the same core. The core is made of a high permeability material such as soft iron and its function is to increase the magnetic flux linking the coils, hence increasing the mutual inductance between each pair of them. When an alternating current is passed through one of the coils, almost all the flux linking it also passes through the other coil. Depending on the number of turns in the coils, the induced e.m.f. may be many times larger or smaller than the e.m.f. driving the transformer. If the induced e.m.f. is now allowed to drive a load, electrical power can be very efficiently transferred from the generator to the load circuit. The transformer thus provides a convenient way of changing the amplitude of an alternating e.m.f.

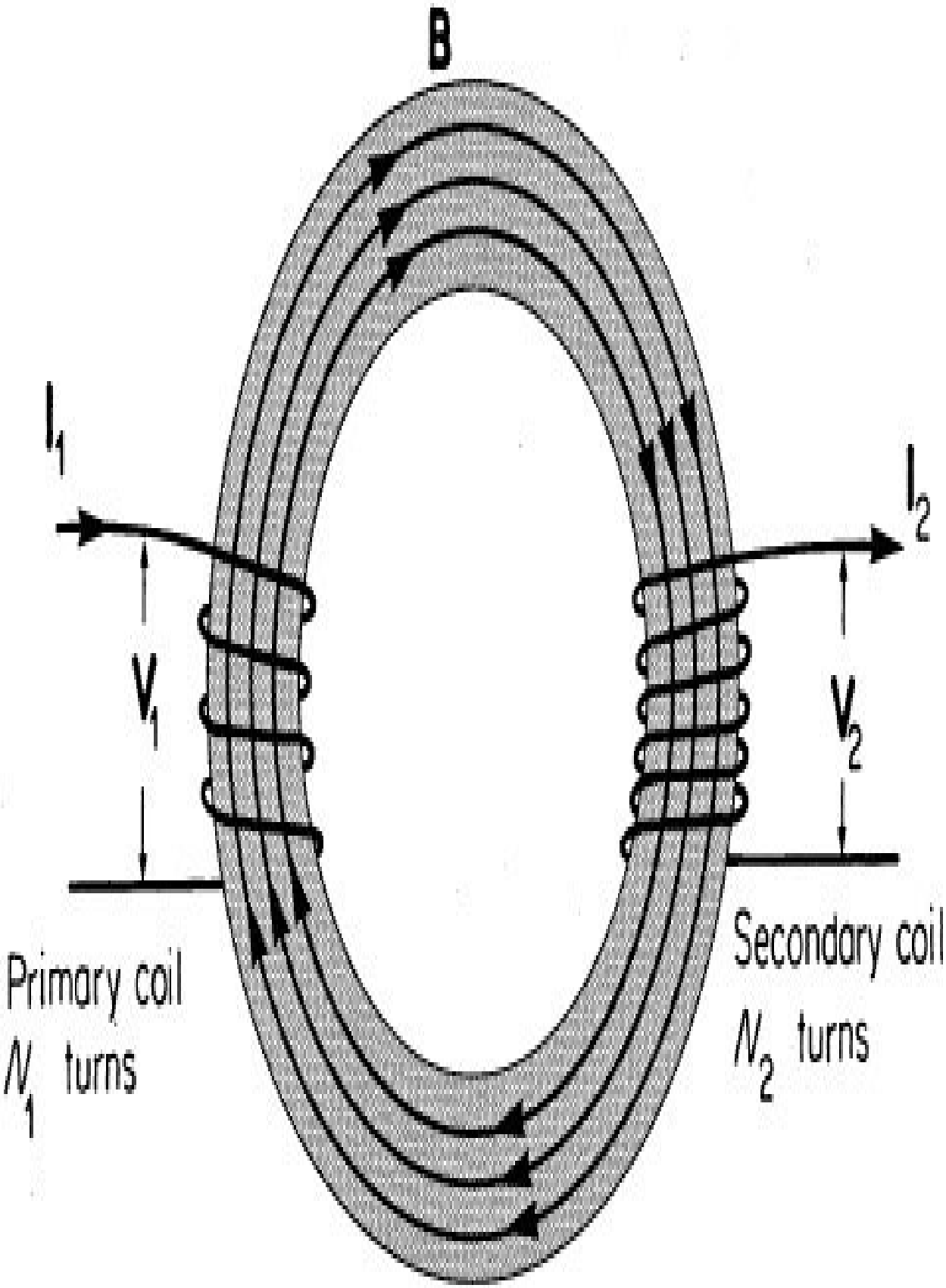
At first sight it may seem surprising that we should think of treating a transformer as a linear component, since the magnetic field in a ferromagnetic material is not linearly related to the current generating the field. However, the magnitudes of the currents in the transformer coils are not very sensitive to the magnetic properties of the core material, provided only that it does have a high permeability. The behaviour of the transformer is in fact determined mainly by the ratio of the numbers of turns in its coils. We can relate the currents and voltages in the coils by approximate expressions in which the turns ratio is the only transformer parameter. These expressions—which are linear—are very simple, and since they illustrate several important properties of transformers it is useful to imagine an ‘ideal’ transformer for which they hold exactly. The ideal transformer is discussed in detail in the next section, some applications of transformers are described in section 8.5.2 and then in section 8.5.3 we shall explain the ways in which real transformers depart from the ideal.

8.5.1 The ideal transformer

An ideal transformer is one which has perfect magnetic coupling between its coils, and which dissipates no energy. The coils must therefore have zero resistance, and no heat losses occur within the core itself. The core of an ideal transformer is assumed to have a very high permeability indeed, yet not to exhibit saturation. This ensures that the self-inductance of each coil is enormous, and also that the core traps all the magnetic flux in the transformer. In the ideal toroidal transformer in Figure 8.21 this is represented by drawing a high density of field lines in the core and none outside. The field lines have neither beginning nor end (because there are no free magnetic poles) and each line makes a complete circuit within the core. All the lines pass through any cross-section of the core—in other words the whole of the magnetic flux is carried by the core and threads both the coils wound on it.

The left-hand coil in Figure 8.21, wound with N1 turns, is connected to an external source of e.m.f. which generates an alternating voltage V1 across its terminals. It is called the primary coil. The other coil—the secondary coil—has AT2 turns and is connected to a load with impedance Z2. By applying Faraday's law to the two coils, we can now find the induced e.m.f. V2 appearing across the terminals of the secondary coil. Exactly the same flux, of magnitude Φ, say, passes through each turn of both coils. An induced e.m.f. ($-d\Phi/dt$) is induced around each turn, and since there is no resistance, the induced e.m.f.s balance the e.m.f.s

Figure 8.21. An ideal transformer.


B
I_1
I_2
V_1
V_2
Primary coil
N_1 turns
Secondary coil
N_2 turns

across the terminals. Thus

(8.24)

and

(8.25)

Eliminating dΦ/dt, it follows immediately that

(8.26)

writing n for the turns ratio N2/N1 The output voltage of the secondary is in phase with the primary voltage, and the ratio of output to input amplitudes is the turns ratio N2/N1

The currents I1 and I2 flowing in the primary and secondary circuits are related by Ampère's law to the line integral of B around a circuit threading the coils. Consider a circuit of length l lying entirely within the core. If the core has a uniform cross-section A, the magnetic field is nearly uniform over any cross-section, and has the average magnitude B = ΦA. In applying Ampère's law to this circuit, we must remember that the secondary current is flowing in such a direction as to oppose the flux changes caused by the primary current. The two currents are therefore flowing in the opposite sense around the lines of B and

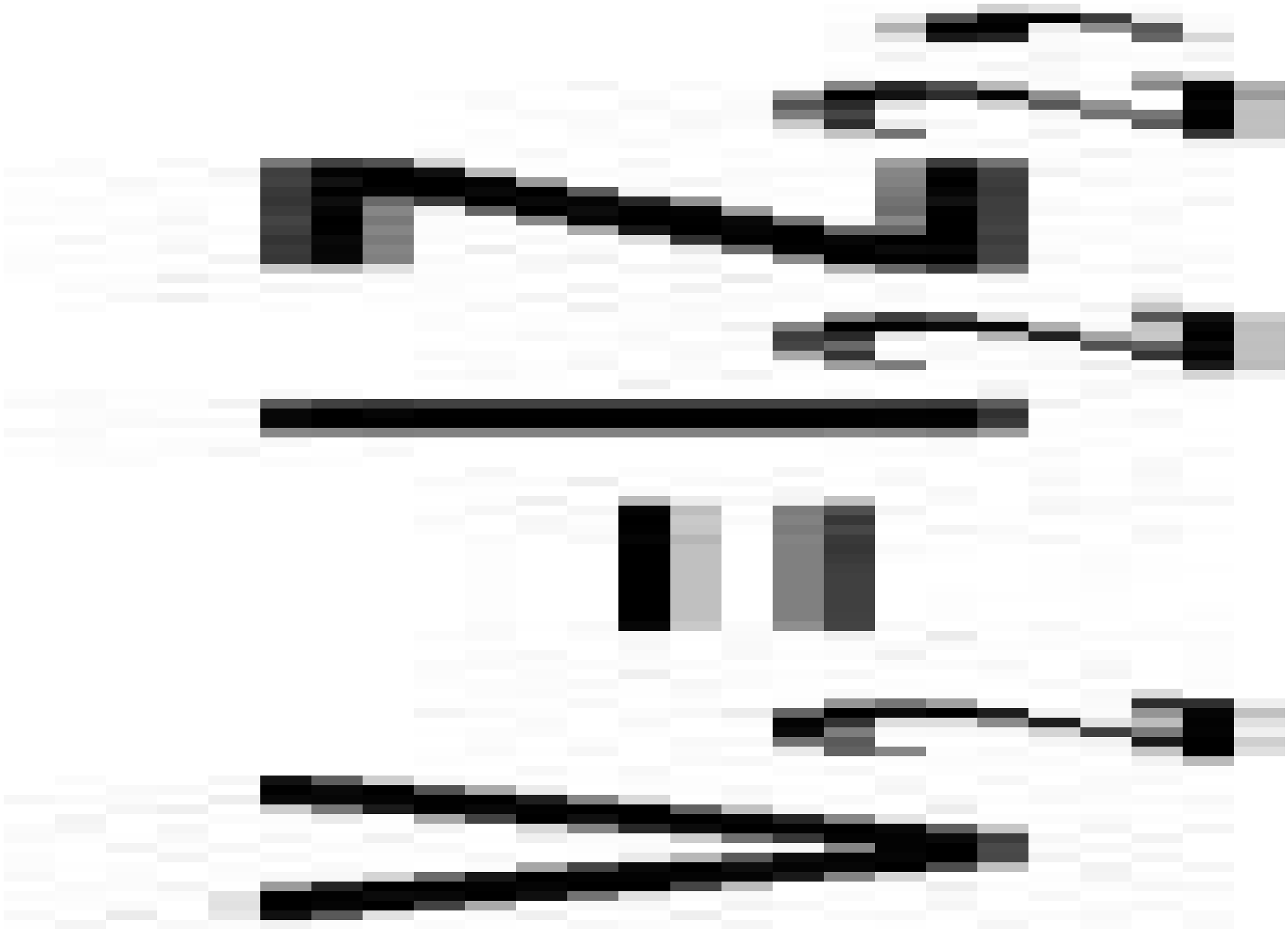
(8.27)

since μ is indefinitely large. Practically no current can flow in the primary when the secondary is open-circuited, because the self-inductance of the primary coil of an ideal transformer is so large. When the transformer is on load, the fluxes generated by the primary and secondary coils almost exactly cancel each other. This is why the transformer behaves as a linear circuit element in spite of the non-linear properties of its ferromagnetic core. Ignoring the small term on the right-hand side of Equation (8.27), we find

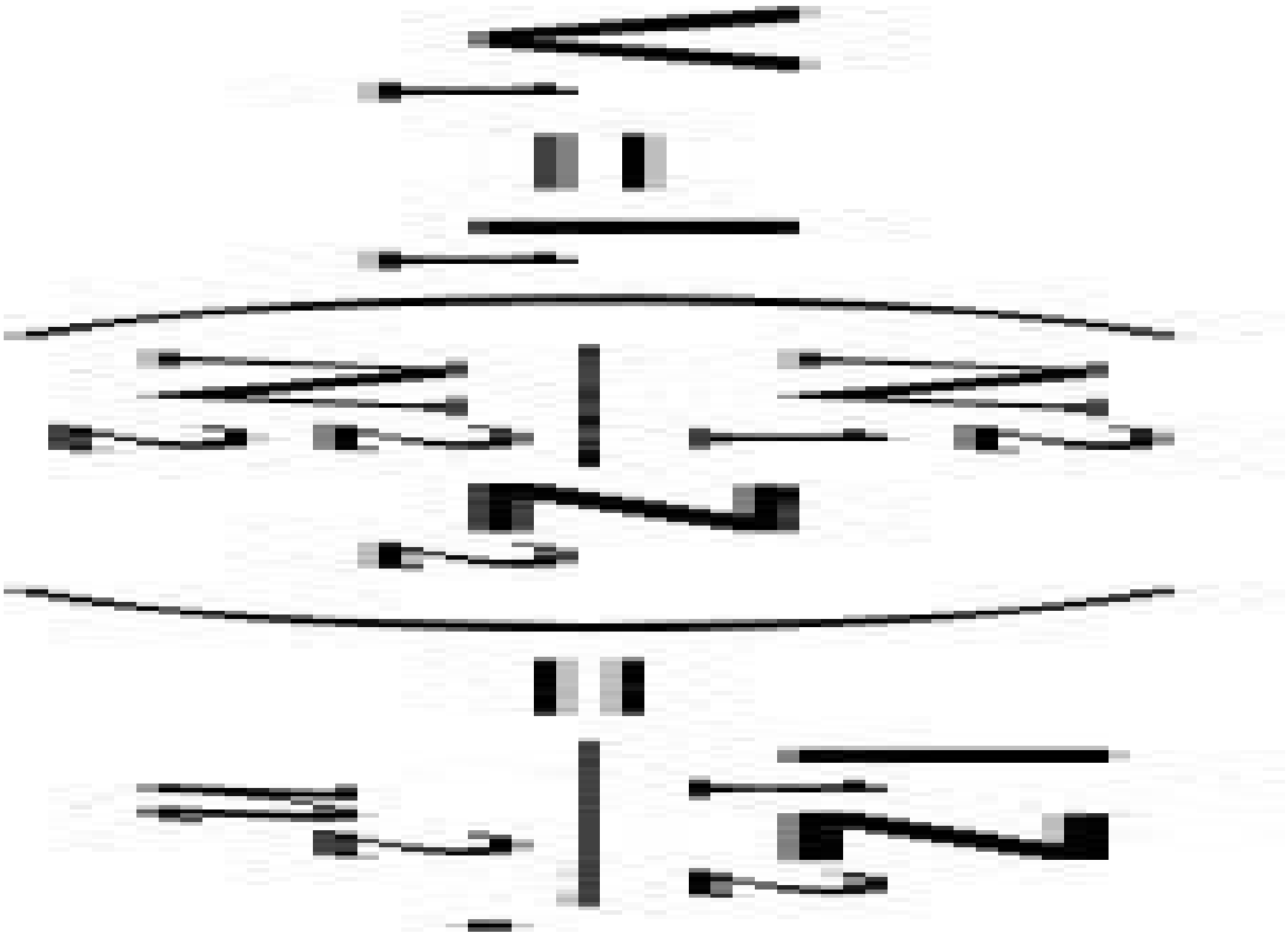
(8.28)

The output and input currents are in phase, and the ratio of their amplitudes is the inverse of the turns ratio N2/N1.

The value of the currents depends on the load impedance. In the secondary circuit,



and hence from Equations (8.26) and (8.28),



To the primary circuit, the transformer looks like an impedance $Z2/n^2$, i.e. it has an input impedance

(8.29)

Similarly, if the output impedance of the circuit driving the primary coil is Z1, the output impedance of the transfomer is

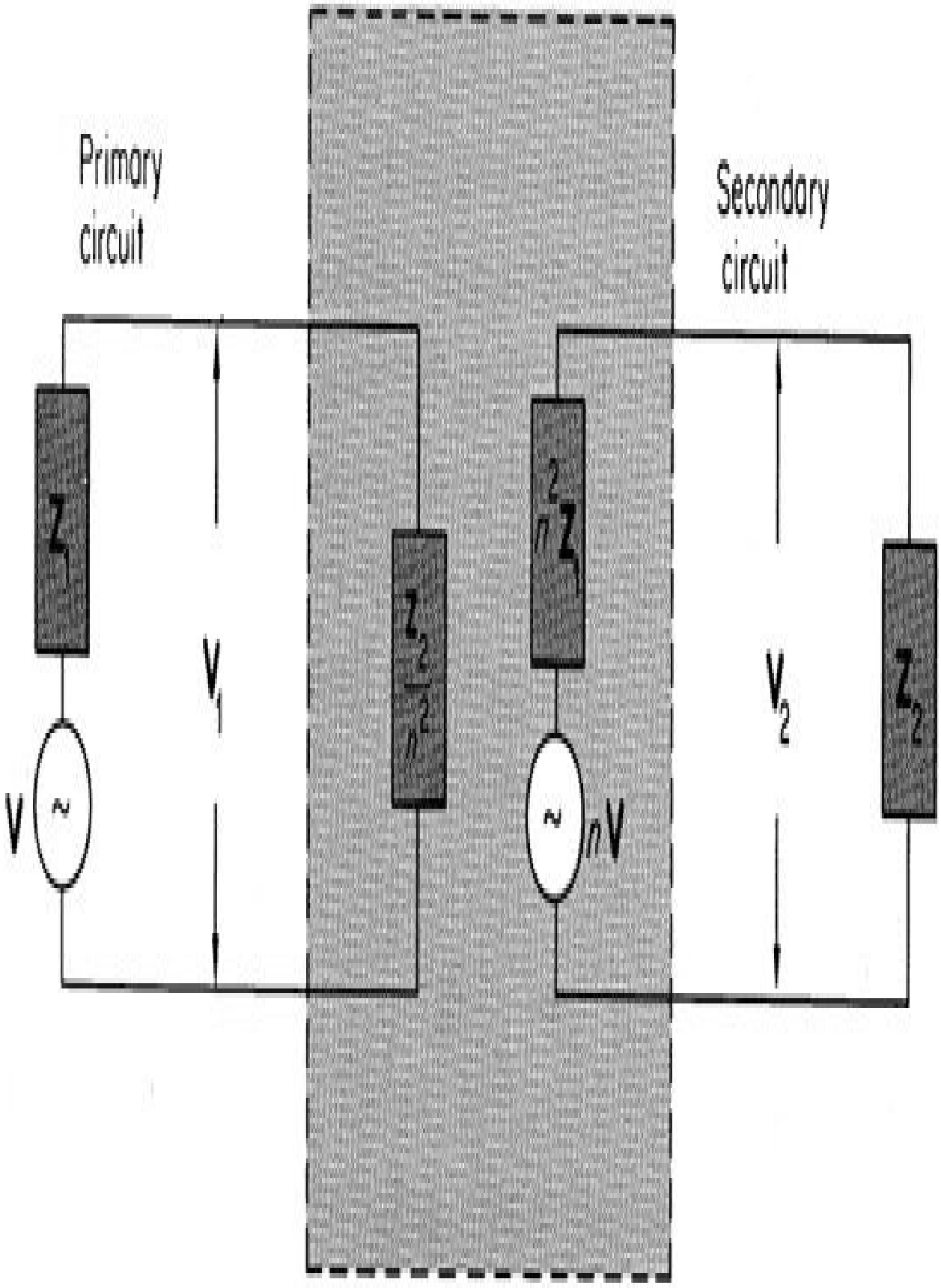
(8.30)

(It is left to the reader to show that Equation (8.30) is correct.)

In the analysis of the primary circuit, the transformer is simply replaced by a component with impedance Zin. In the secondary it must be represented by source of e.m.f. in series with the impedance Zout. This the equivalent circuit for an ideal transformer driven by a generator of voltage V and internal impedance Z1 is as shown in Figure 8.22.

Note that since the ideal transformer has no losses, all the power delivered to the primary coil is transferred to the secondary circuit. The primary and secondary voltages are in phase with one another, and so are the primary and secondary currents. The power factors in the two circuits are the same, with a value $\cos\theta$ which is determined by the nature of the load. The power output of the secondary is

Figure 8.22. The equivalent circuit for an ideal transformer.


Primary
circuit
Secondary
circuit
Z_1
V_1
$\frac{Z_2}{n^2}$
V
~
$n^2 Z_1$
~
nV
V_2
Z_2

$$V_2 I_2 \cos\theta = (nV_1)\left(\frac{I_1}{n}\right)\cos\theta$$

$$= V_1 I_1 \cos\theta$$

= power delivered to the primary.

8.5.2 Applications of transformers

The equations we have derived for an ideal transformer are good approximations to the behaviour of a real one. The transformer is aptly named, since we have found that it will pass electrical power from one circuit to another, while transforming voltage, current and impedance. In this section some applications of transformers are briefly described to illustrate the usefulness of each of their transformations.

(i) Voltage transformation. This is the most common and obvious application of transformers. Electrical power is normally obtained from a main supply at 110 or 240 volts, and in any circuit for which the mains voltage is unsuitable, a transformer is usually to be found. For example, a mains transformer will be used to drive the rectifier in a high-voltage D.C. power supply. This transformer requires more turns in the secondary coil than in the primary, and since it increases the voltage, is referred to as a step-up transformer. Similarly a step-down transformer is used if a low voltage supply is wanted for an electronic circuit. Although the turns ratio in the step-down transformer is small, it does not follow that the secondary coil is much smaller than the primary. Power is dissipated in the windings and in the core of a real transformer: the power loss is small compared with the power transferred from primary to secondary, but the maximum permissible temperature rise of the transformer will nevertheless limit the amount of power it can transfer. In a step-down transformer the secondary has to carry a higher current than the primary, and must be wound with thicker wire to reduce the resistive heating. The current density, and hence the power density, are the same in the primary and secondary coils if they contain the same volume of conductor. Whenever transformers are to handle appreciable amounts of power, it is most economical to have equal heat losses in the primary and secondary coils, which must therefore be made about the same size.

(ii) Current transformation. Electrical power is transmitted over large distances by conductors carried high above the ground on their ungainly pylons. The conductors must be so high because they operate at a high voltage—400 kV for

the British grid—and the main reason for choosing a high voltage is tomake the current as small as possible for a given power, in order to minimize the resistive losses in the conductors. At generating stations, huge amounts of power are converted to low current and high voltage by transformers, and it is obviously important that the transformer losses should be small. In this respect big transformers approach quite closely to the ideal, with losses of only about 0.2% of the input power.

(iii) Impedance transformation. There are many kinds of circuit which will only function efficiently if they are connected to a load whose impedance has a particular value. For example, the delay lines described in section 8.4.2 transmit all signals containing frequencies below cut-off only if they are matched by a terminating load equal to their low-frequency characteristic impedance. It may happen that the impedance of the load is such that it is impracticable to construct a delay line to match it. Thus if the load impedance were 1 Ω, the inductance per section of a matching line with a cut-off at 1.5 MHz would have to be a small fraction of a microhenry, less than the stray inductances always present in the circuit. In this case a step-down transformer with a turns ratio of 1:10 would convert the load impedance to 100Ω, and match it to the line designed in section 8.4.2.

Impedances also may need to be matched in circuits designed to transfer power. If an A.C. generator of R.M.S. voltage V and output impedance R1 (purely resistive) is driving a load resistance R2, then the R.M.S. current in the circuit is $I = V/(R1 + R2)$. The power delivered to the load is $I^2R2 = V^2R2/(R1 + R2)^2$, which has its maximum value at R1 = R2, when the same power is dissipated in the generator and the load.

Isolating transformers

Another useful property of the transformer is that the primary and secondary circuits may be completely insulated from one another. An isolating transformer can therefore transfer A.C. signals between two circuits which are at different

D.C. levels. If the coils of an ideal isolating transformer have the same number of turns, it will change the D.C. level without affecting the A.C. signal at all; voltage, current and impedances are all unaltered.

8.5.3 Real transformers

None of the characteristics assumed for an ideal transformer can be achieved in practice. In a real transformer

(i) the magnetic coupling between the coils is not perfect;

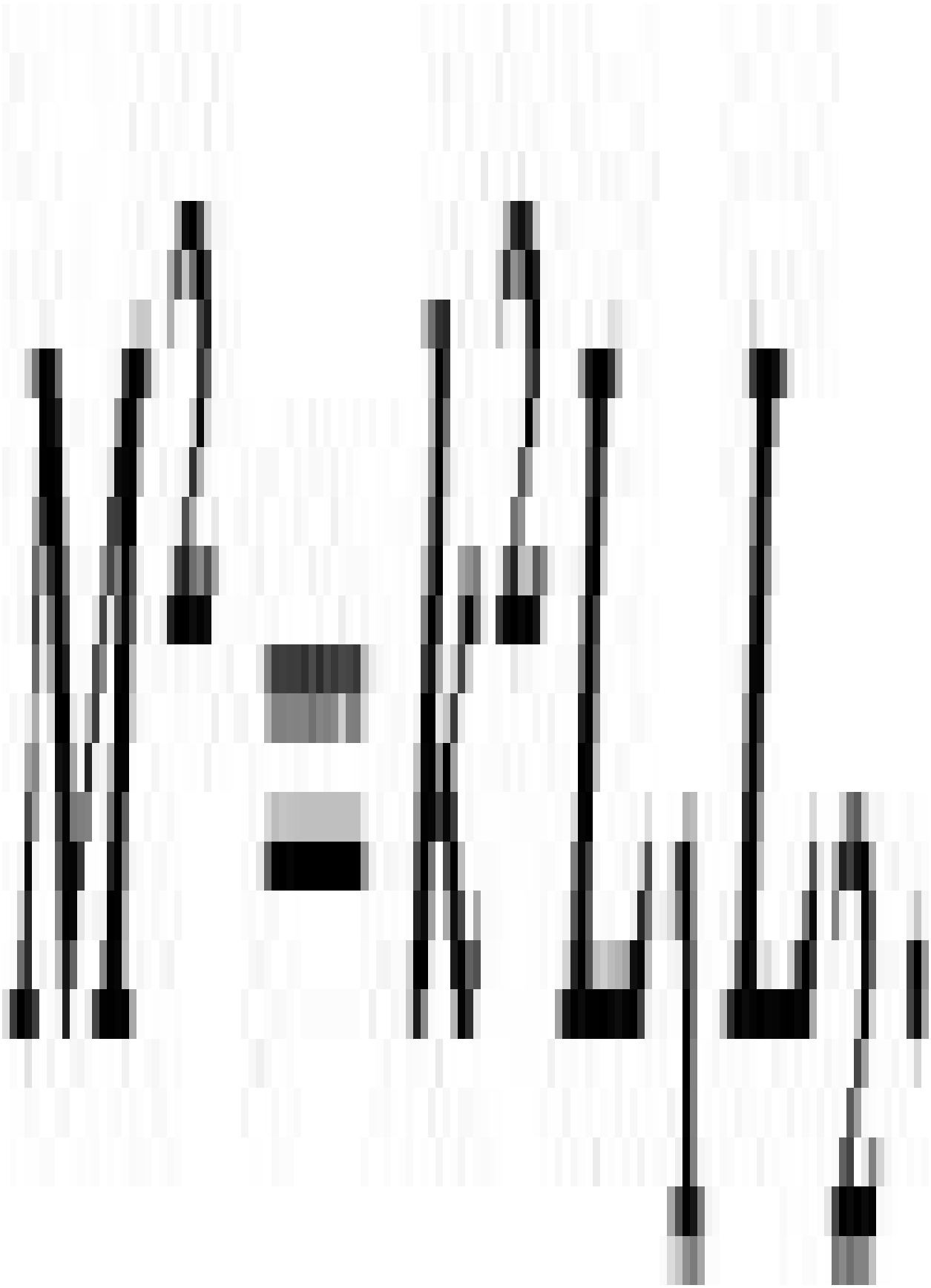
(ii) there are power losses in the coils windings and in the core;

(iii) at high magnetic fields the core material exhibits saturation.

The way in which each of these factors affects the performance and design of transformers will be considered in turn.

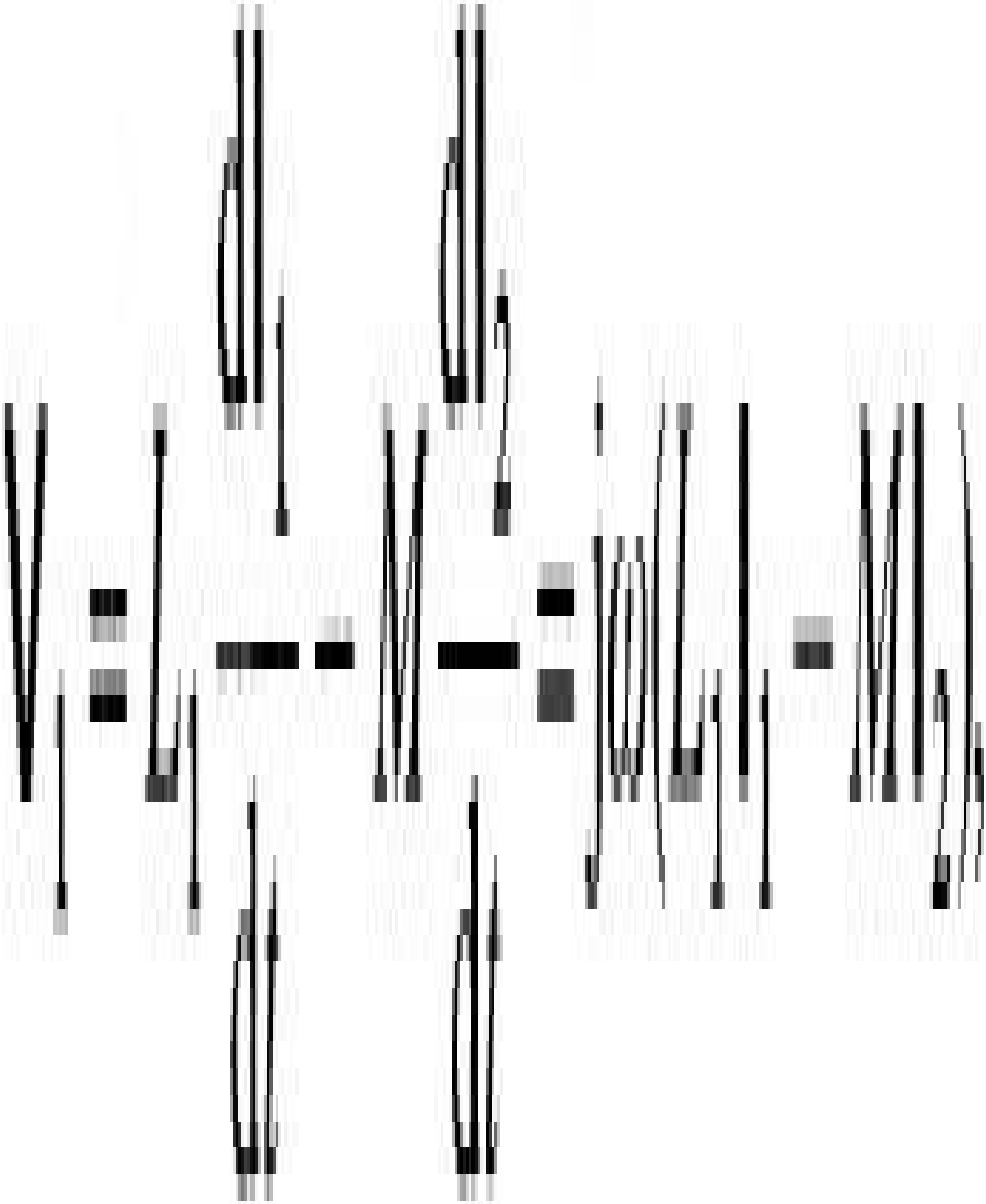
Leakage flux

In spite of the high permeability of the soft iron used in transformer cores, the flux linkage between the coils is never quite complete. The self-inductances L1 and L2 of the primary and secondary coils are related to their mutual inductance M by Equation (6.25):



Here the coupling coefficient k represents the fraction of the flux through one of the coils which threads the other. The remaining fraction (1 – k) of the flux does not link the coils and does not contribute to the e.m.f. induced in the secondary. To find what effect this leakage flux has, we shall rewrite the transformer equations in terms of the self- and mutual-inductances, still assuming that the coils have zero resistance. The current in the secondary flows so as to oppose the flux changes caused by the primary, and for the angular frequency ω the primary and secondary voltages are

(8.31)

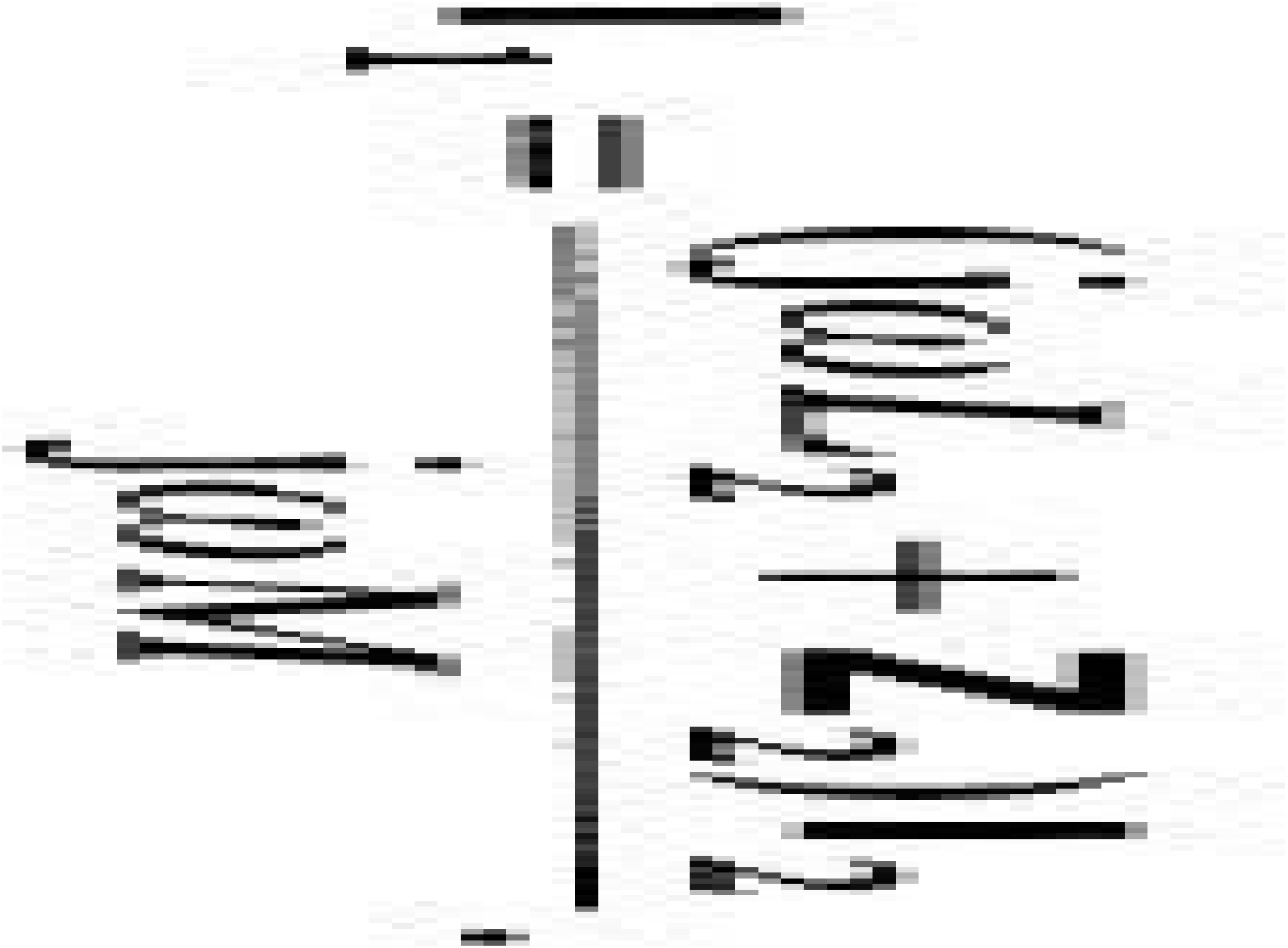


and

(8.32)

when the secondary is driving a load Z2.

From Equation (8.32),



Substituting for I2 in Equation (8.31), we find the input impedance of the transformer to be

$$\mathbf{Z}_{in} \equiv \frac{\mathbf{V}_1}{\mathbf{I}_1} = j\omega L_1 - \frac{(j\omega M)^2}{j\omega L_2 + \mathbf{Z}_2}$$

$$= j\omega L_1 \left(1 - \frac{M^2}{L_1 L_2}\right) + \frac{j\omega \mathbf{Z}_2 M^2 / L_2}{j\omega L_2 + \mathbf{Z}_2}$$

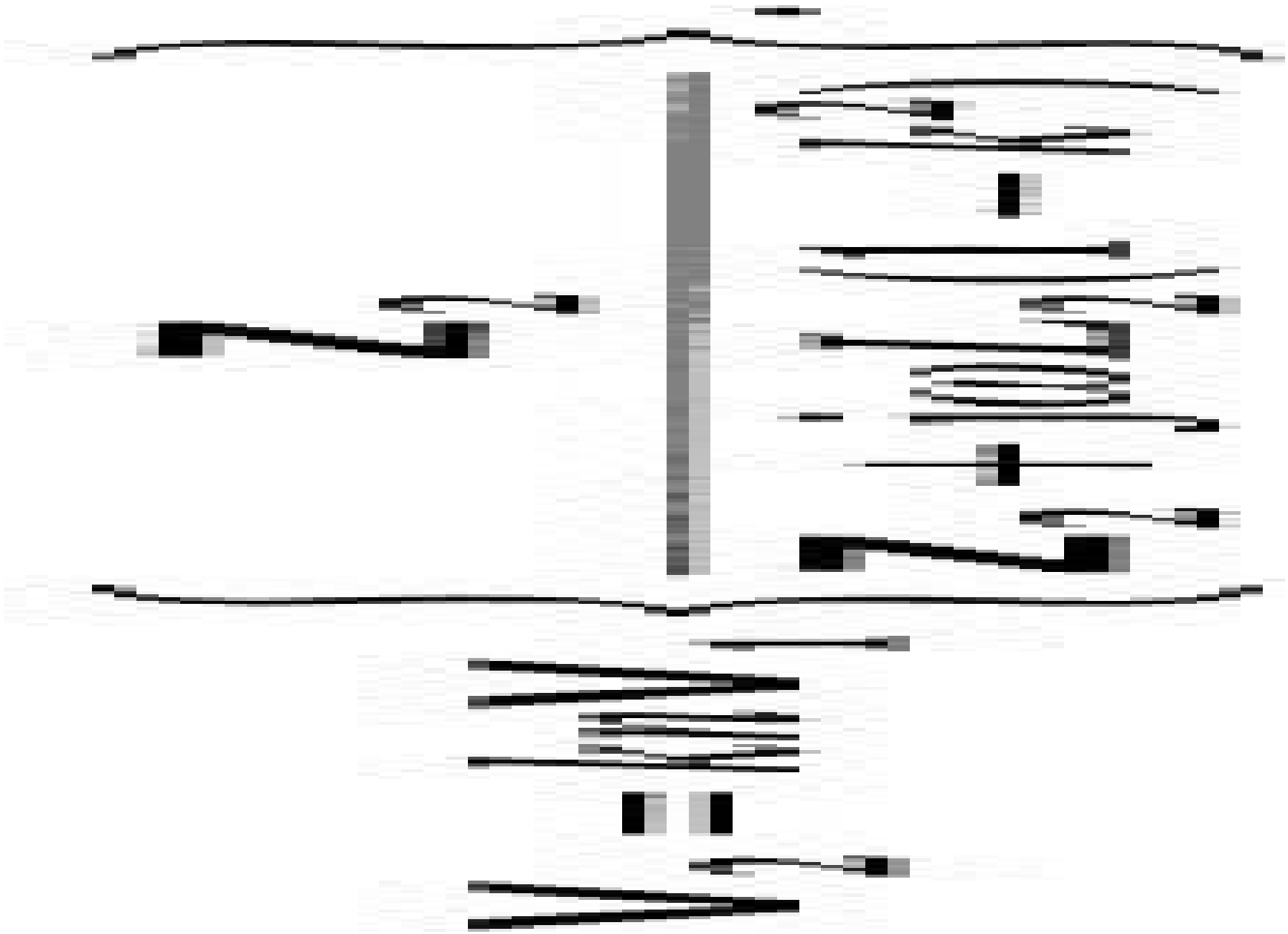
$$= j\omega L_1 (1 - k^2) + \frac{j\omega k^2 \mathbf{Z}_2 L_1}{j\omega L_2 + \mathbf{Z}_2}.$$

In the second term we now approximate by putting $L2 = n^2L1$. This approximation leads to little error in practice. The input impedance becomes

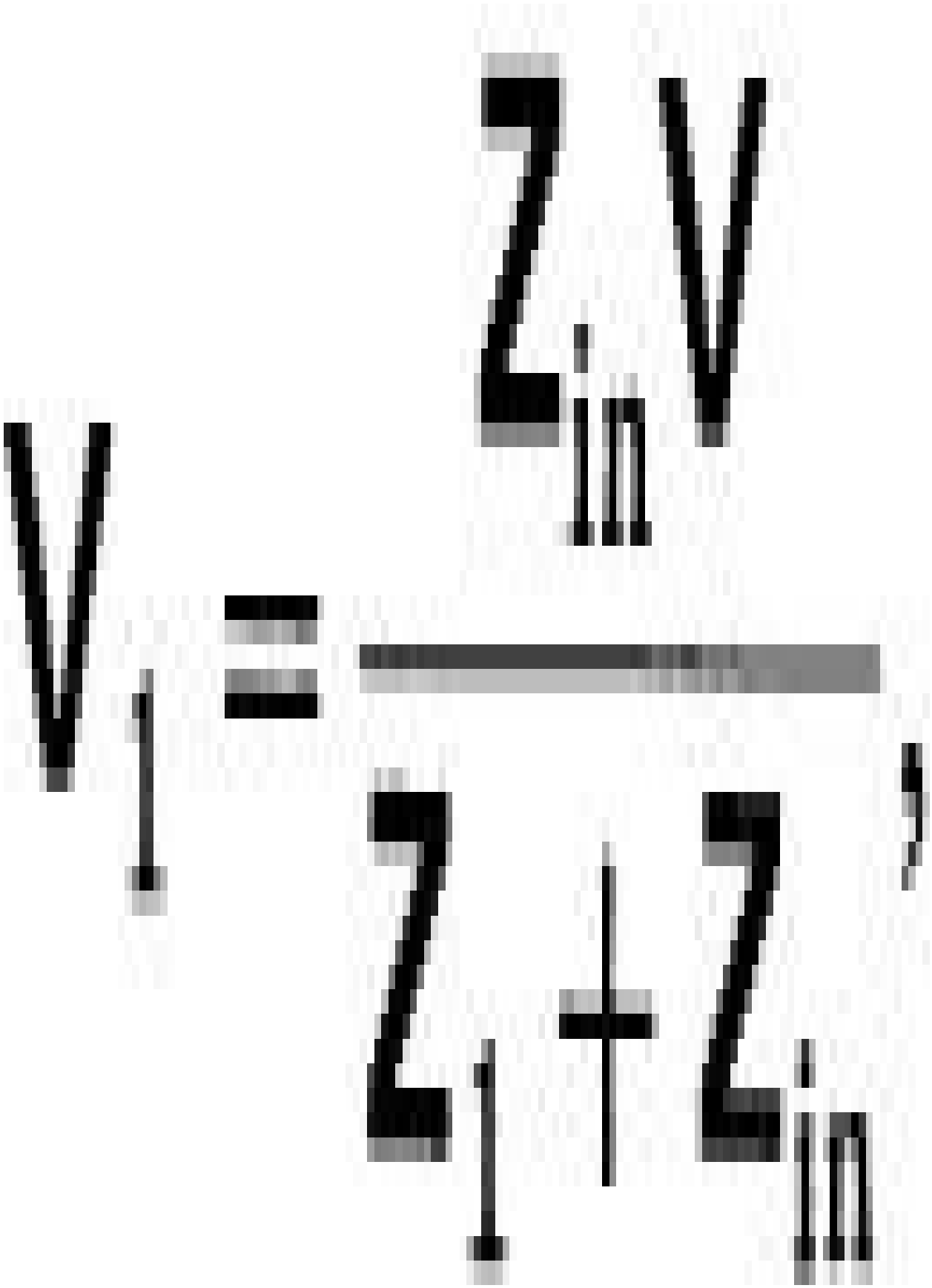
(8.33)

the same as the input impedance of the equivalent circuit shown in Figure 8.23.

By eliminating I1 and I2 from Equations (8.31) and (8.32), the secondary voltage is found to be

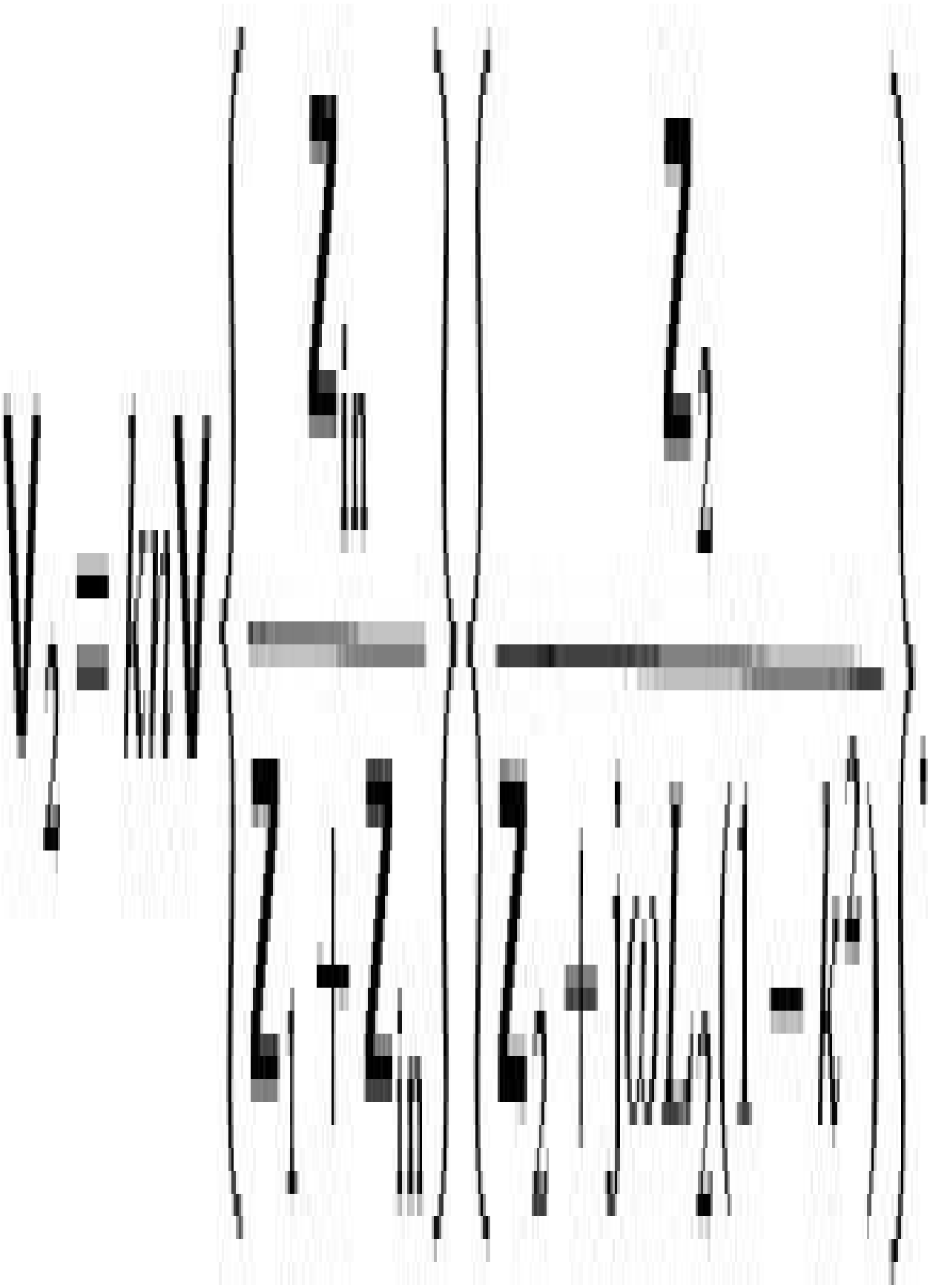


On open circuit the output voltage is knV1 smaller by the factor k than for an ideal transformer. The input voltage V1 itself depends on the output impedance Z1 of the generator driving the transformer, and if this generator develops an open circuit voltage V, then



and

(8.34)



The equivalent circuit which has this output voltage is shown on the right-hand side of Figure 8.23. As one would expect, the primary and secondary circuits look very similar. Their impedances are the same except that the labels 1 and 2 are interchanged, and in the secondary circuit the turns ratio n is inverted. The coupling coefficient k must be very near to unity, so that the series inductance $L2(l - k^2)$ in the output circuit is small compared to the load impedance Z2. High coupling is often achieved by winding the coils on top of one another, as in the transformer sketched in Figure 8.24, which has a core carrying flux symmetrically around both sides of the coils.

Figure 8.23. Equivalent circuit of a transformer including leakage flux.

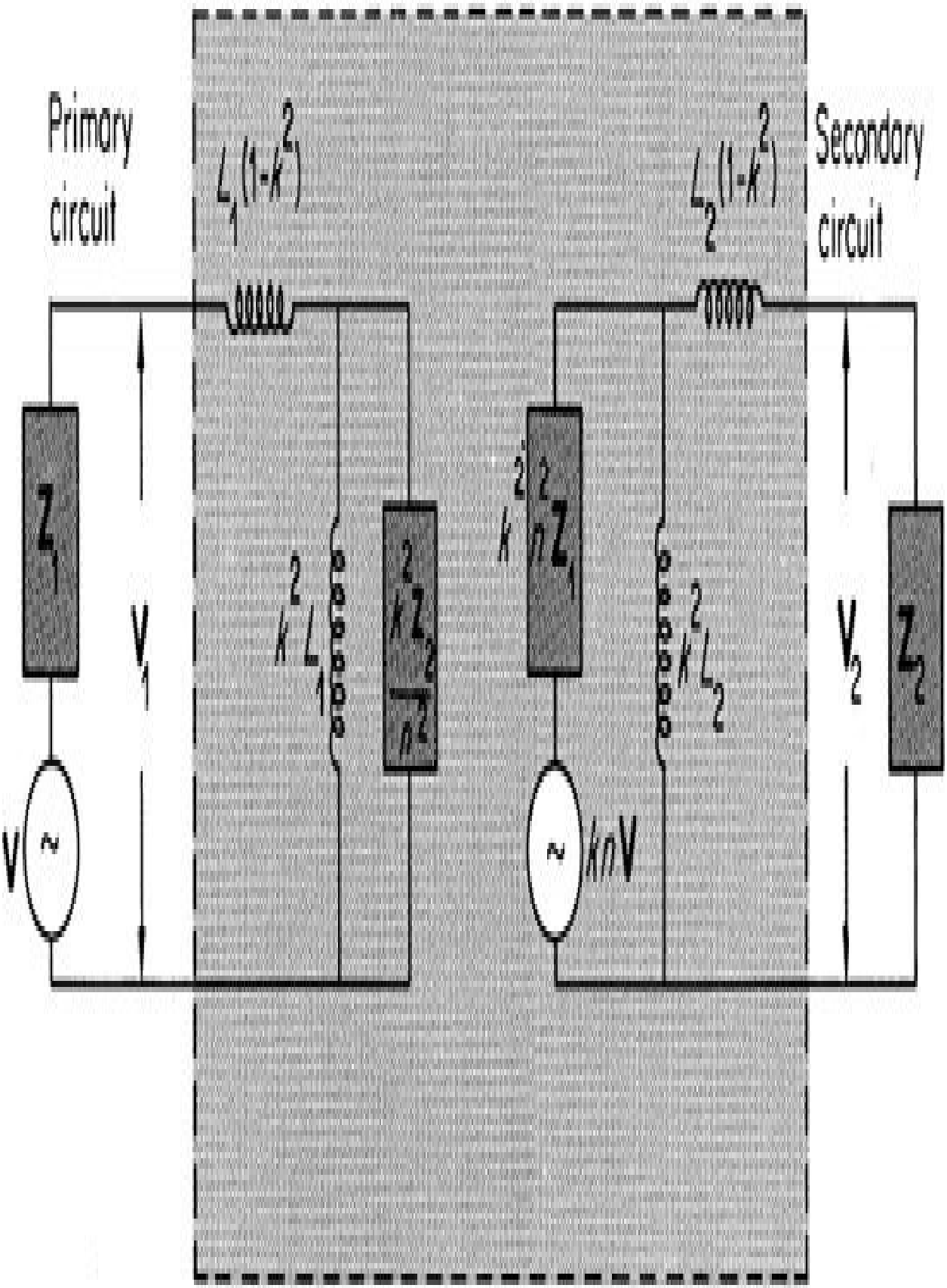
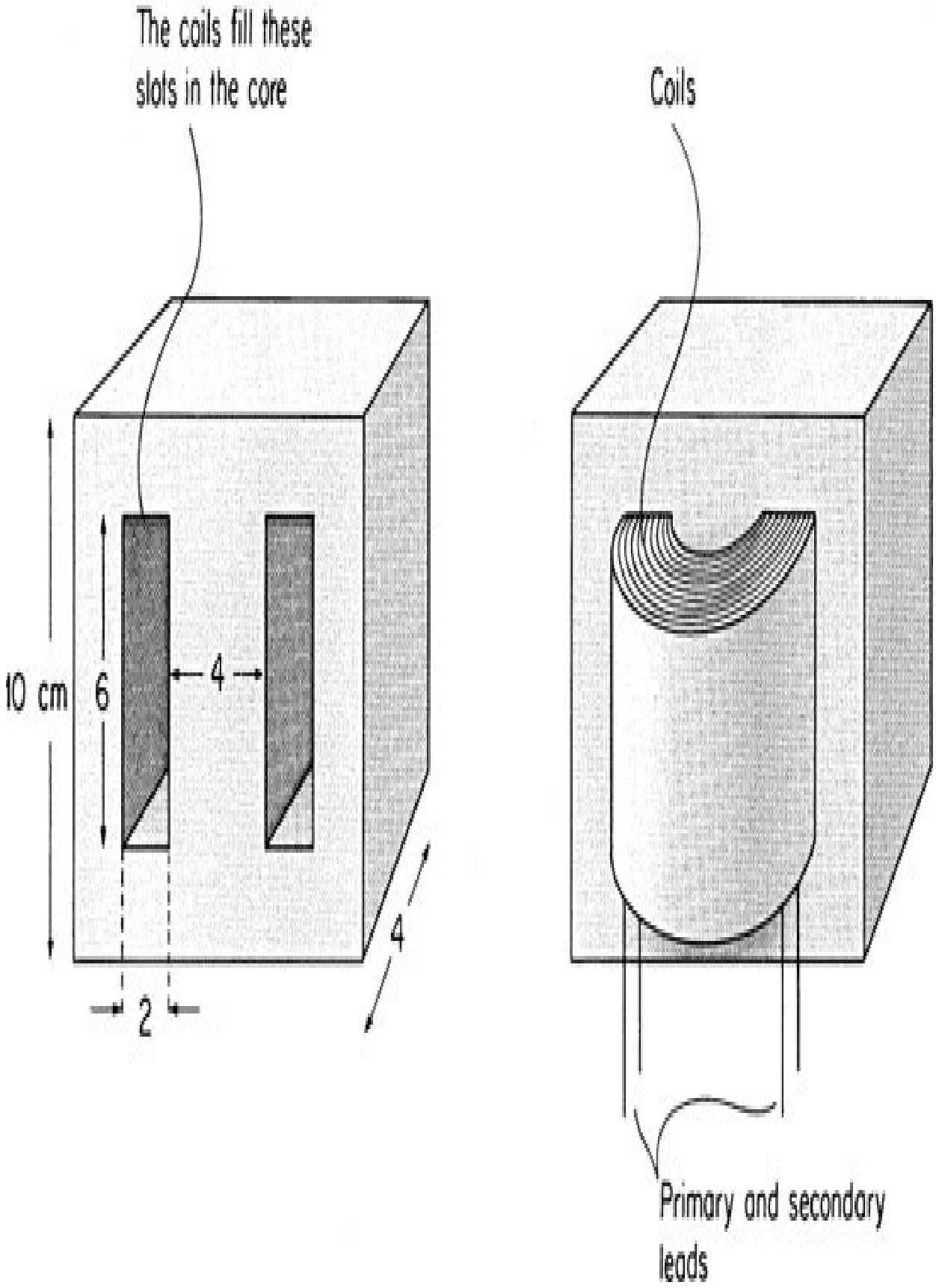

Primary
circuit
$L_1(1-k^2)$
$L_2(1-k^2)$
Secondary
circuit
Z_1
V_1
$k^2 L_1$
$\frac{k^2 Z_2}{n^2}$
$k^2 n^2 Z_1$
$k^2 L_2$
V_2
Z_2
V
~
~
knV

Figure 8.24. Transformer coils are wound together on the same part of the core in order to minimize leakage flux.

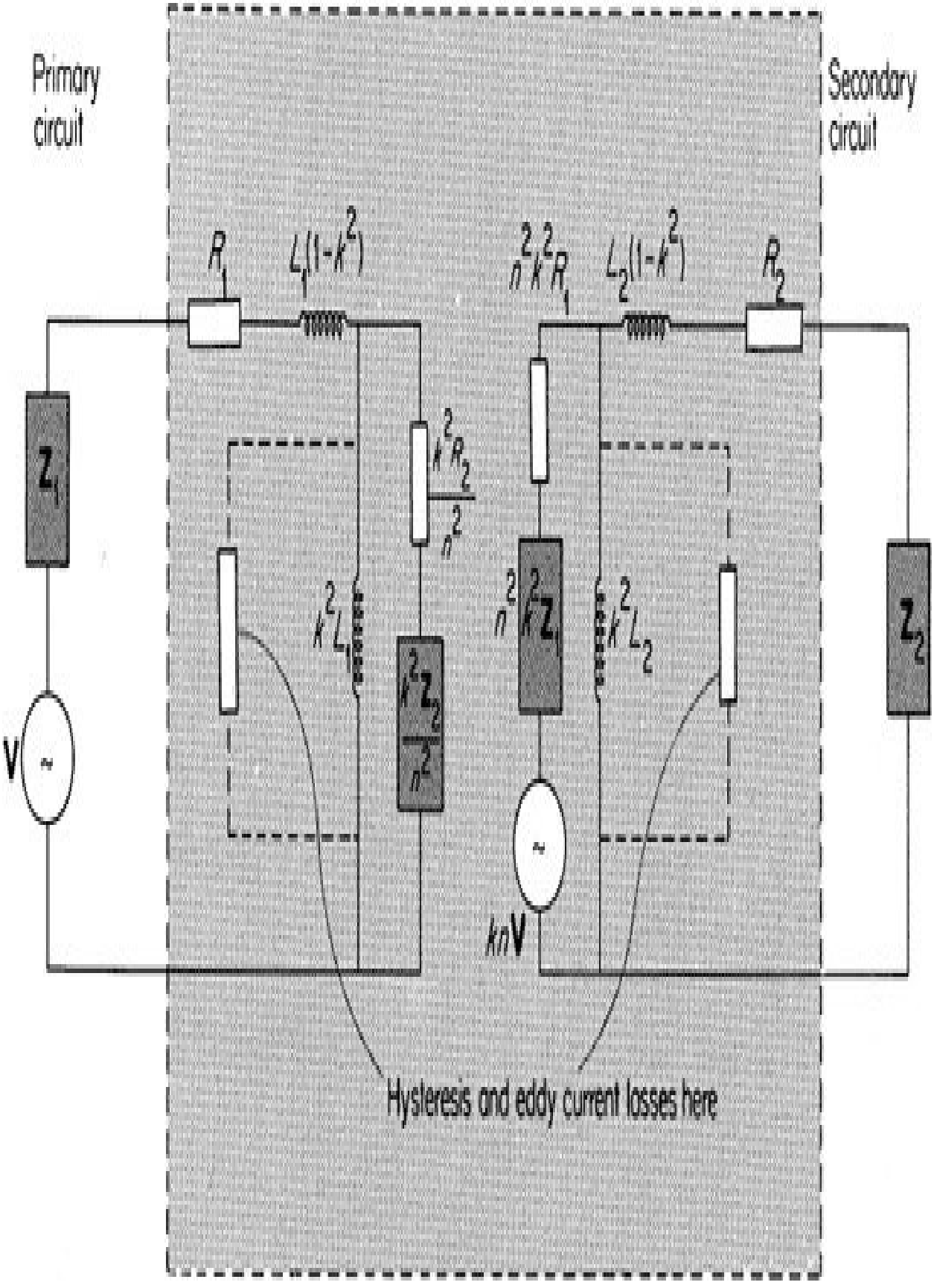

The coils fill these slots in the core
Coils
10 cm
6
4
4
2
Primary and secondary leads

Losses

(a) Coil resistance. The resistance of the coils is easily included in the equivalent circuit by adding series resistors R1 and R2 as in Figure 8.25. The primary resistance alters the output impedance of the transformer, requiring a resistor n^2k^2R1 to be included in the secondary: similarly a resistor k^2R2/n^2appears in the primary circuit. These resistors affect the low-frequency performance of the transformer. Even on open-circuit, the output voltage of the transformer falls unless the angular frequency $\omega \gg R1/L1$ and $\omega \gg R2/L2$. The resistance per unit inductance of the coils of an audio-frequency transformer is typically about 20 ohms/henry, leading to a frequency $R1/2\pi L1$ of about 3 Hz. At 3 Hz the performance is very poor, but a good audio transformer will operate at 30 Hz with a power loss of no more than 3 dB.

(b) Hysteresis losses. Once in every cycle of the alternating current, the iron of the transformer core is taken around its hysteresis curve (see section 6.3.5). An amount of energy proportional to the area of the hysteresis curve is dissipated during each cycle. The loss is not simply proportional to the voltage, since the shape of the hysteresis curve changes as the maximum flux is altered. However, at a fixed frequency the hysteresis loss can be roughly represented in the equivalent circuit by a resistor in parallel with the transformer inductance.

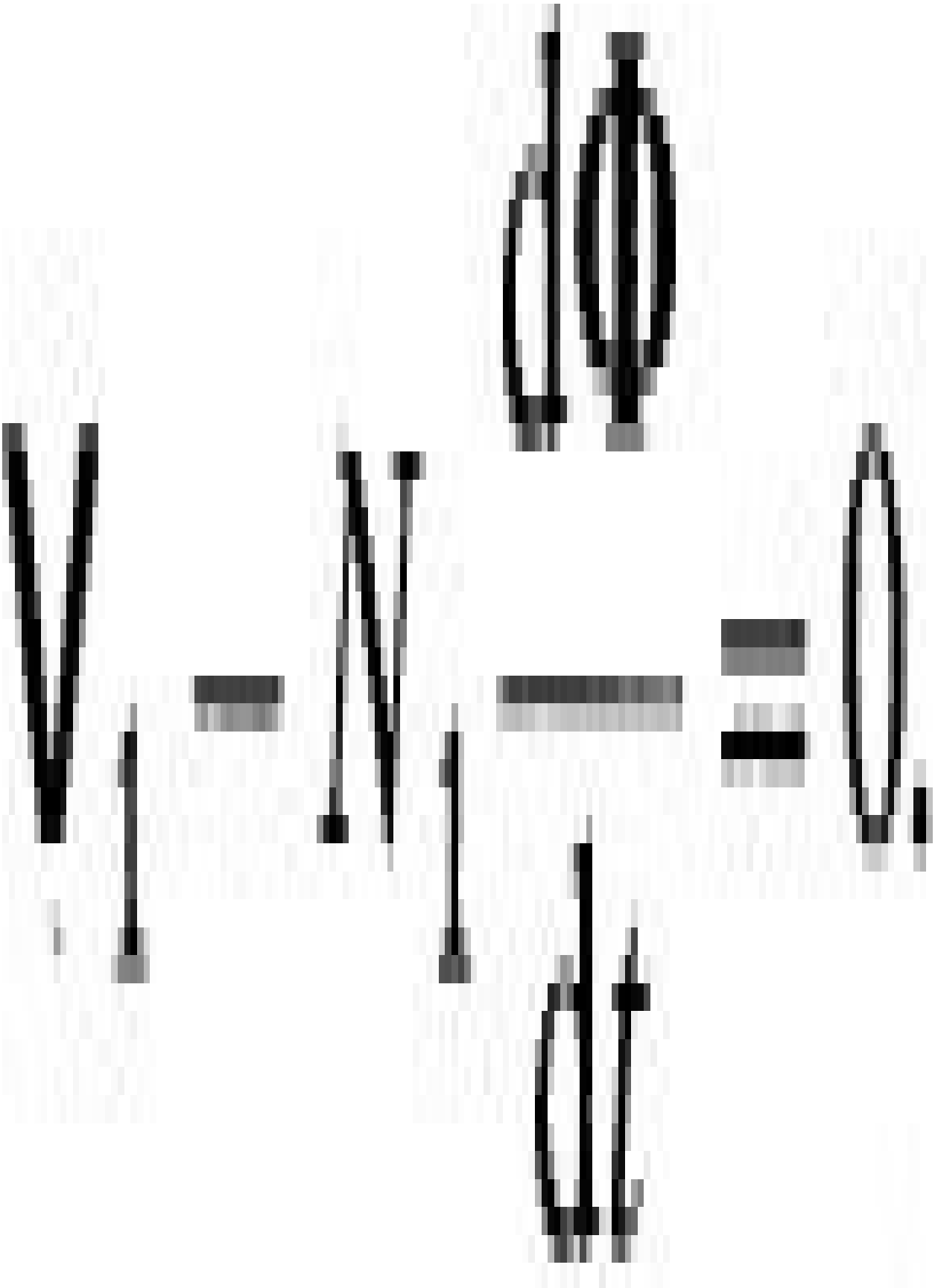
Figure 8.25. Equivalent circuit for a transformer, including losses.


Primary circuit
Secondary circuit
R_1
$L_1(1-k^2)$
$n^2k^2R_1$
$L_2(1-k^2)$
R_2
Z_1
$\frac{k^2R_2}{n^2}$
k^2L_1
$\frac{k^2Z_2}{n^2}$
$n^2k^2Z_1$
k^2L_2
Z_2
V
~
~
knV
Hysteresis and eddy current losses here

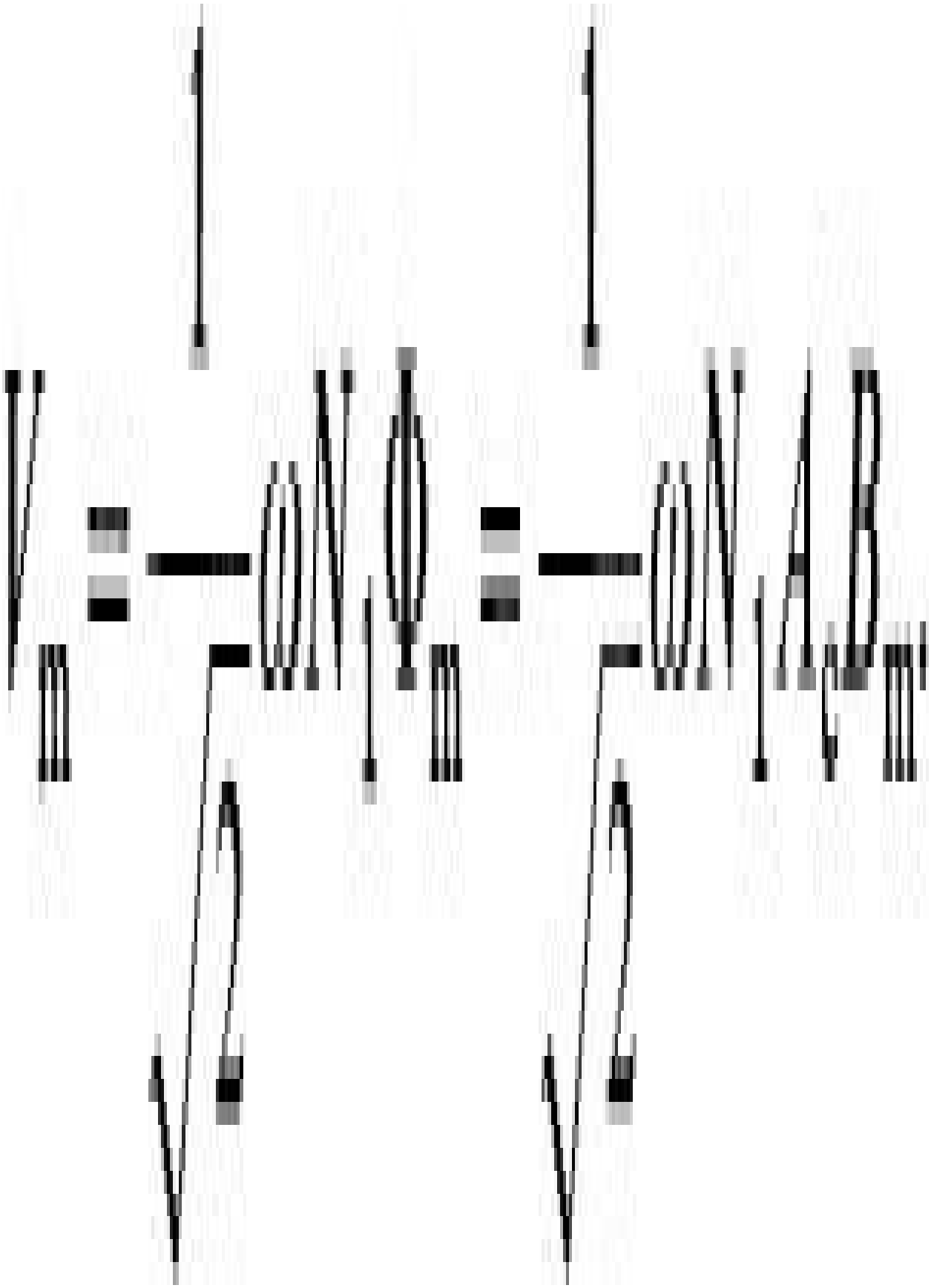
(c) Eddy currents. There are induced e.m.f.s in the transformer core, which cause currents—called eddy currents, first introduced in section 6.1.3—to flow within the core, leading to a further dissipation of energy. The eddy current losses can be reduced by dividing the core into thin laminations which are insulated from one another, and stacked so as to stop the flow of current. Small currents still circulate inside each lamination; for a fixed flux amplitude the rate of change of flux, and hence the eddy current density, is proportional to the angular frequency ω. The power losses thus vary as ω^2, and become increasingly important at high frequency. For very high frequency operation it is not practicable to make laminations sufficiently thin, and cores are then made of powdered iron or of ferrite, a material which has a much higher resistivity than iron. Small-signal transformers with ferrite cores can be used at frequencies up to about 1000 MHz.

Saturation

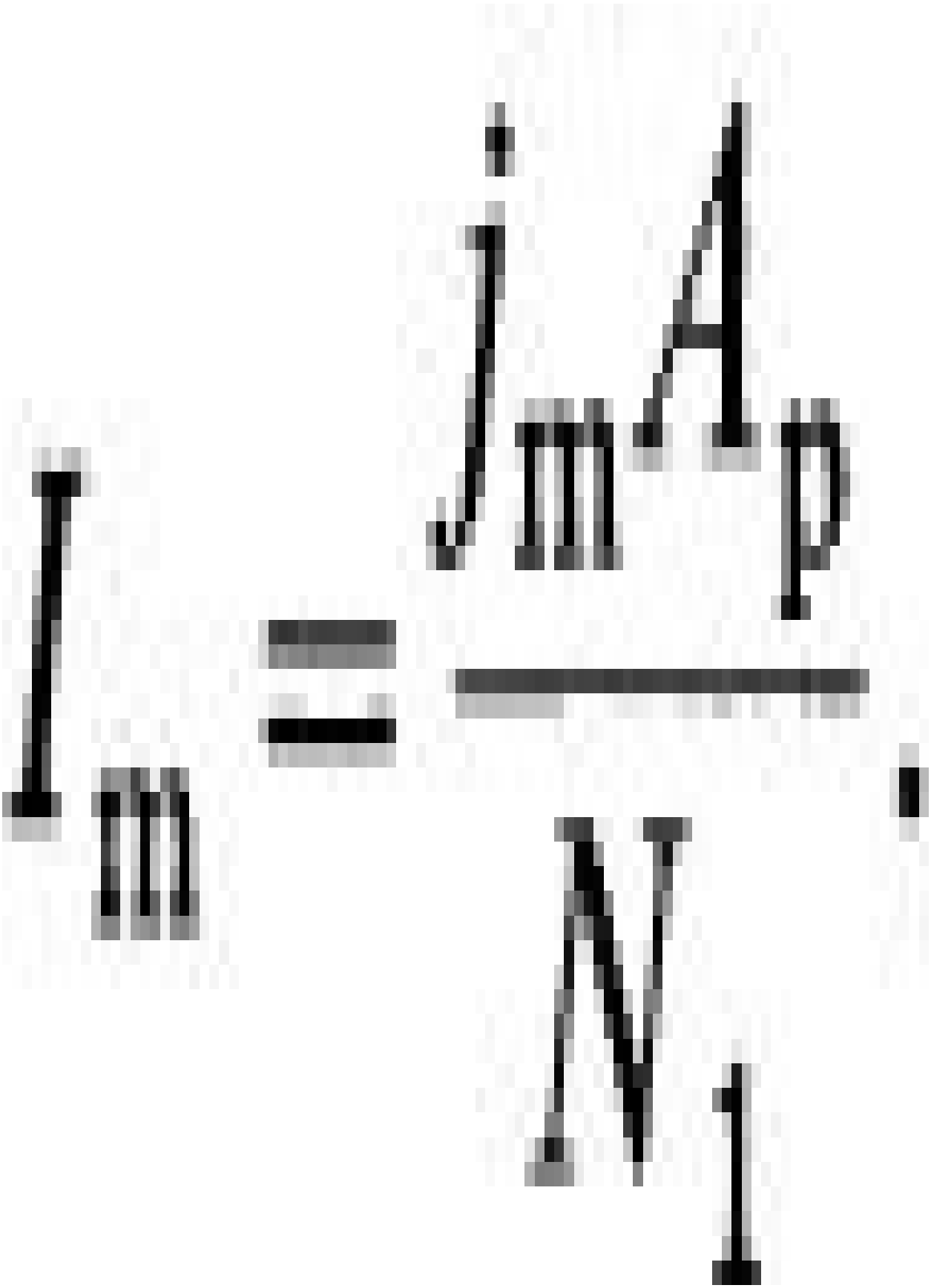
So far we have assumed that the self- and mutual-inductances of the coils in a transformer are constant, independent of the currents they are carrying. This assumption breaks down completely when the ferromagnetic core approaches saturation. The incremental permeability then falls sharply, resulting in smaller inductances and poorer coupling between the primary and secondary coils. A transformer therefore only operates at high efficiency when carrying currents below those which saturate the core. Suppose that Bm is the maximum field which is to be allowed in a particular transformer core. If the core has a cross-sectional area Ac, then the maximum flux is $\Phi m = AcBm$. The flux is related to the voltage V1 on the primary through Equation (8.24):



For an angular frequency ω, the maximum permissible R.M.S. value of the primary voltage is therefore



For a given core size, the value of Vm can be increased by increasing the number of turns on the primary, at the expense of increasing the coil resistance and hence its power dissipation. The power dissipation is limited by the temperature rise which can be tolerated, and in practice this leads to the stipulation that the current density in the windings should not exceed some value jm. If the total area of conductor in a cross-section of the primary coil is Ap, then the maximum current it can carry is

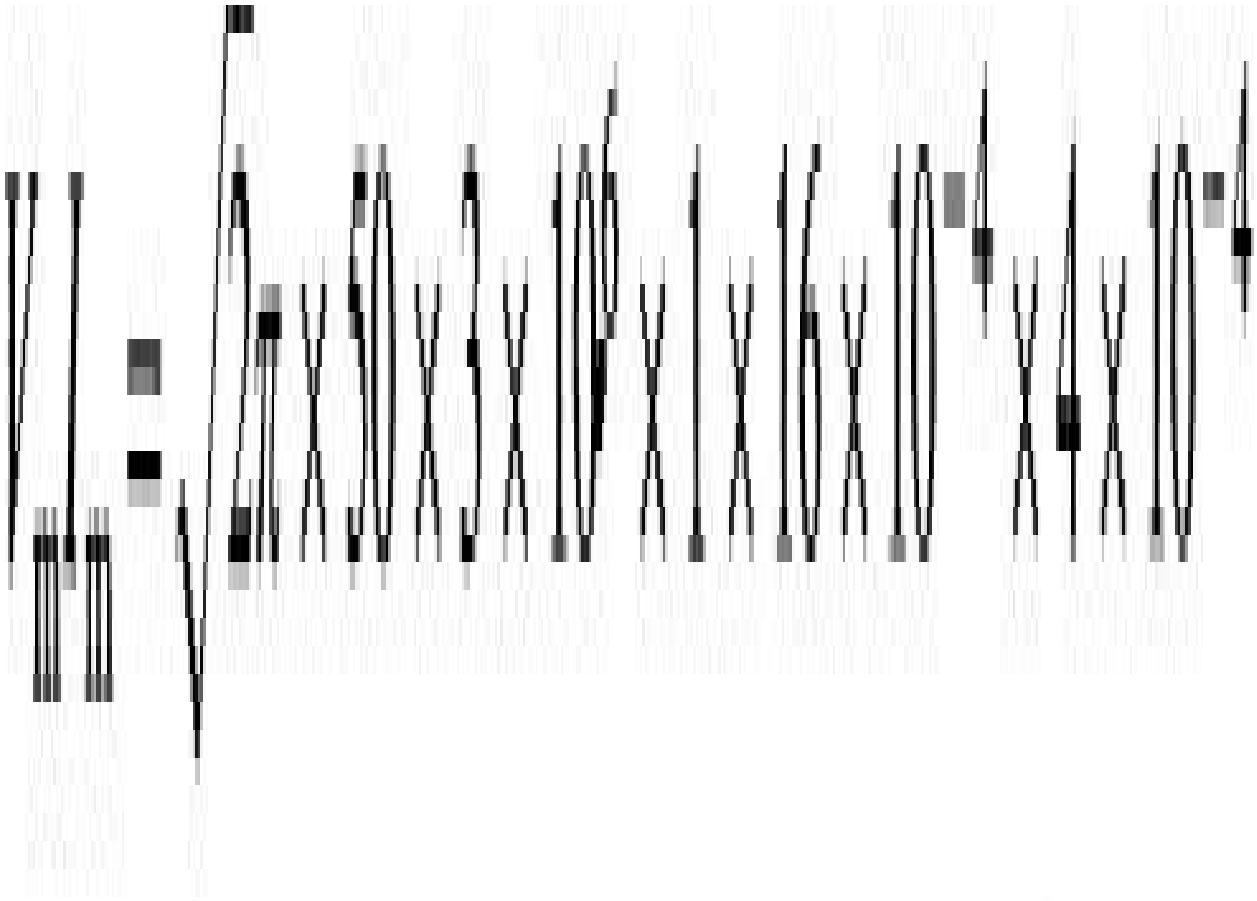


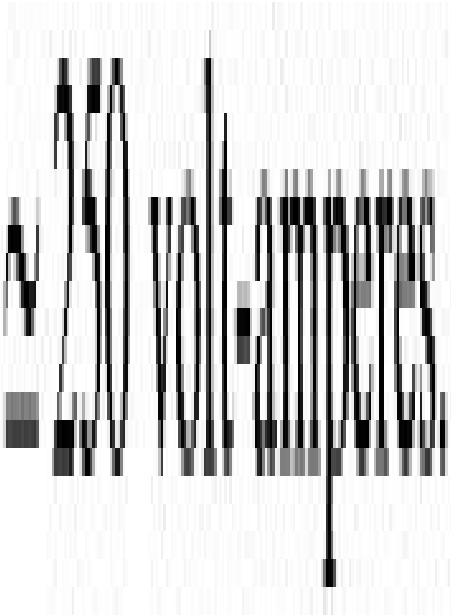
Hence

(8.35)

The same argument applies to the secondary coil, and since the VI products are almost equal in primary and secondary, the secondary coil should have about the same area of conductor as the primary.

Equation (8.35) relates the size of a transformer to its power rating VmIm. We shall illustrate this by working out the rating of a 50 Hz transformer wound on the core shown in Figure 8.24. The core is made of iron with the hysteresis curve drawn in Figure 6.10, and it is to be operated within the minor hysteresis loop in the figure, so that Bm ~ 1T. We shall choose jm to be 3 amps/mm², a typical value for transformers about this size. Allowing a third of the area of the window in the transformer core to be taken up by insulation, Ap = 4 cm², and substituting in Equation (8.35),





(The rating of a transformer is always quoted in volt-amperes instead of watts.)

The resistivity ρ of copper is $1.7 \times 10^{-8}\Omega m$, and the power density in the windings at full current is ρ 0.15 watts/cm^3. Assuming rectangular winding of the coils, the total volume of copper is 128 cm^3, leading to a maximum power dissipation of 20 watts. This is much bigger than the hysteresis losses, which from the area of the hysteresis loop in Figure 6.11 are 3×10^{-3} watts/cm^3 at 50 Hz, or just over one watt for this transformer. The efficiency of the transformer when operated at its maximum rating is thus a little over 90%, a typical value for this power range. Higher efficiencies can be achieved where necessary by increasing the size of the transformer, and in generating stations handling thousands of megawatts of power, the transformers have efficiencies of about 99.8%.

PROBLEMS 8

8.1 Twelve resistors, each of resistance one ohm, are joined to form the edges of a cube. What is the resistance between opposite corners of the cube? (First decide which of the resistors are equivalent when a current is flowing between the two corners, and then apply Kirchhof's rules.)

8.2 Use loop analysis to find the currents I1 and I2 and hence the amplitude of the alternating voltage across the capacitor C2 in Figure 8.26, if C1 C2 and C3 are all 1 μF. The alternating voltages $V_0\cos\omega t$ and $2V_0\cos\omega t$ are in the direction of the arrows.

Figure 8.26.

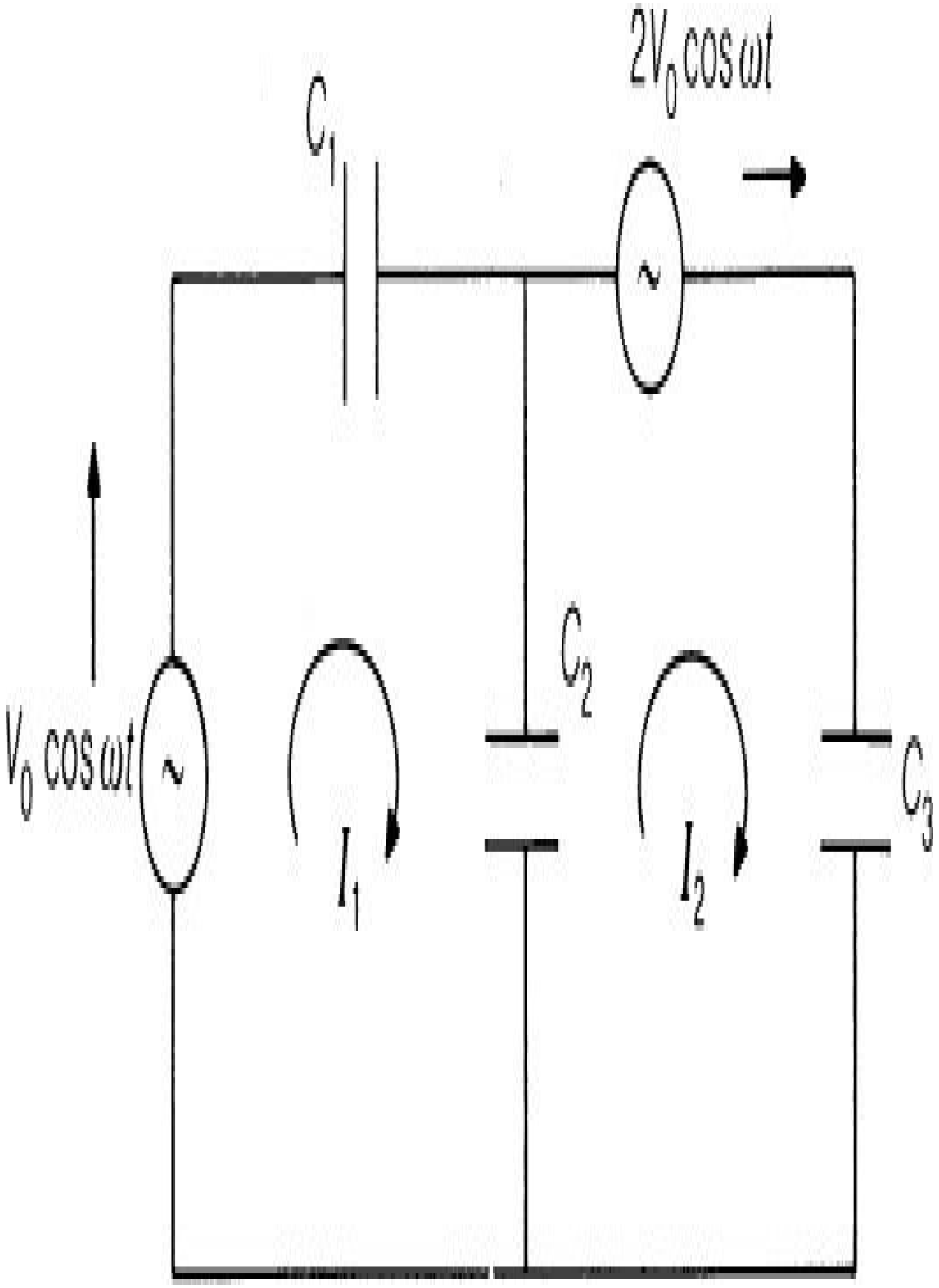
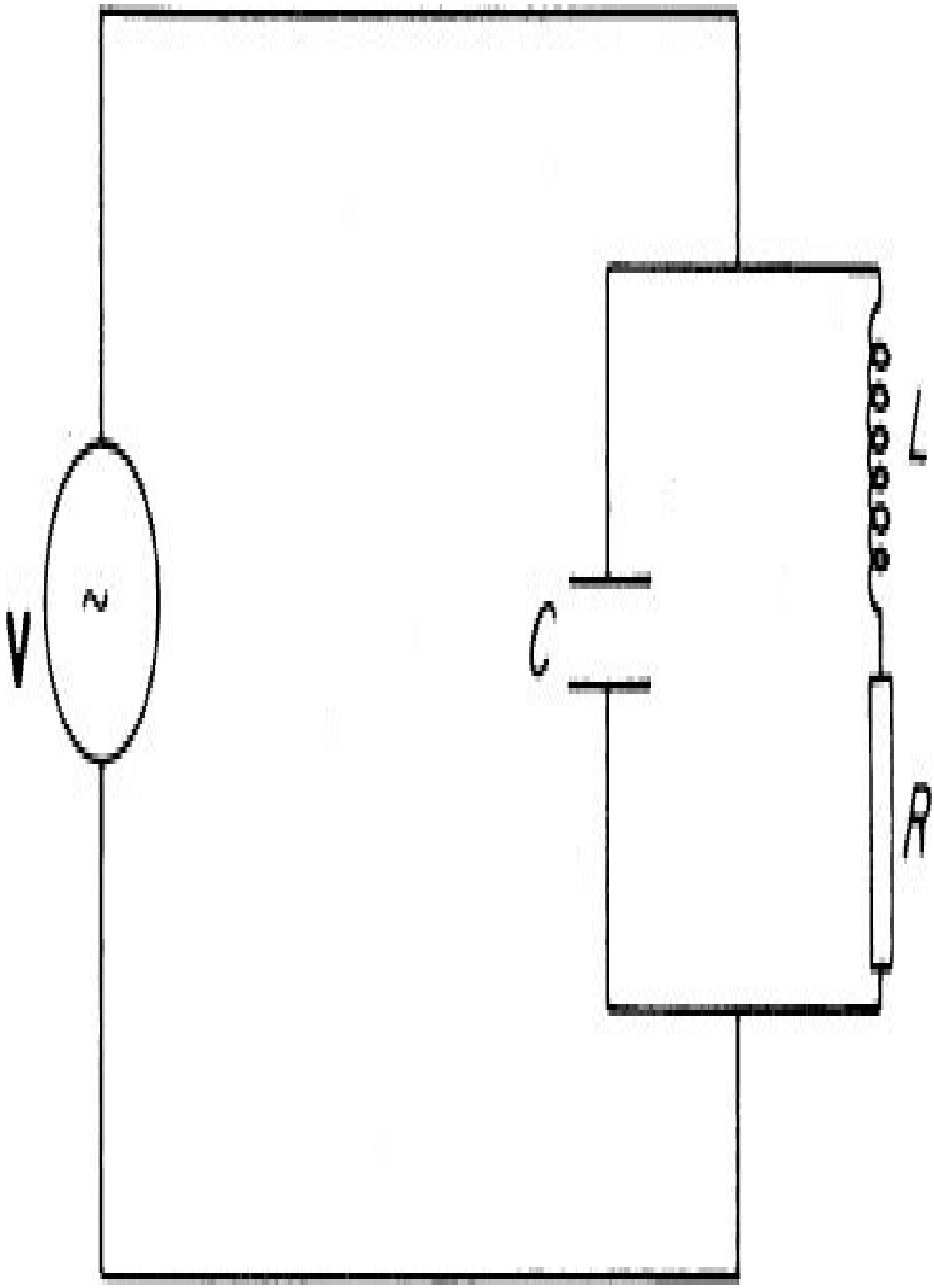

$2V_0 \cos \omega t$
C_1
$V_0 \cos \omega t$
C_2
C_3
I_1
I_2

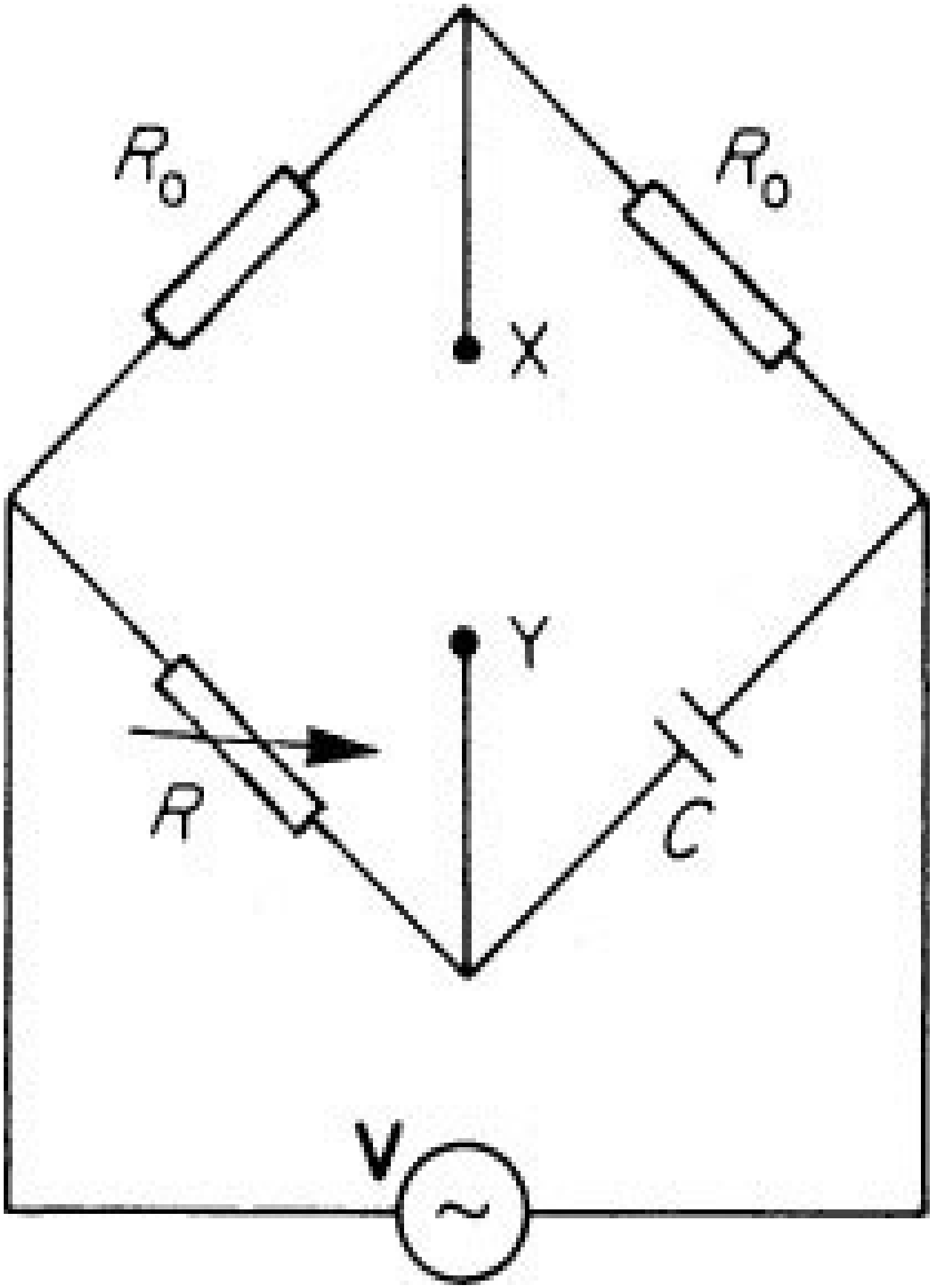
Figure 8.27.


V
~
C
L
R

8.3 The circuit in Figure 8.27 represents a coil and a capacitor placed in parallel across an alternating voltage V. Show that the circuit exhibits resonance and that if L » CR^2 the amplitude of the current drawn from the source of e.m.f. goes through a minimum value near the angular frequency $\omega_0 = 1/$. What is Q for this circuit?

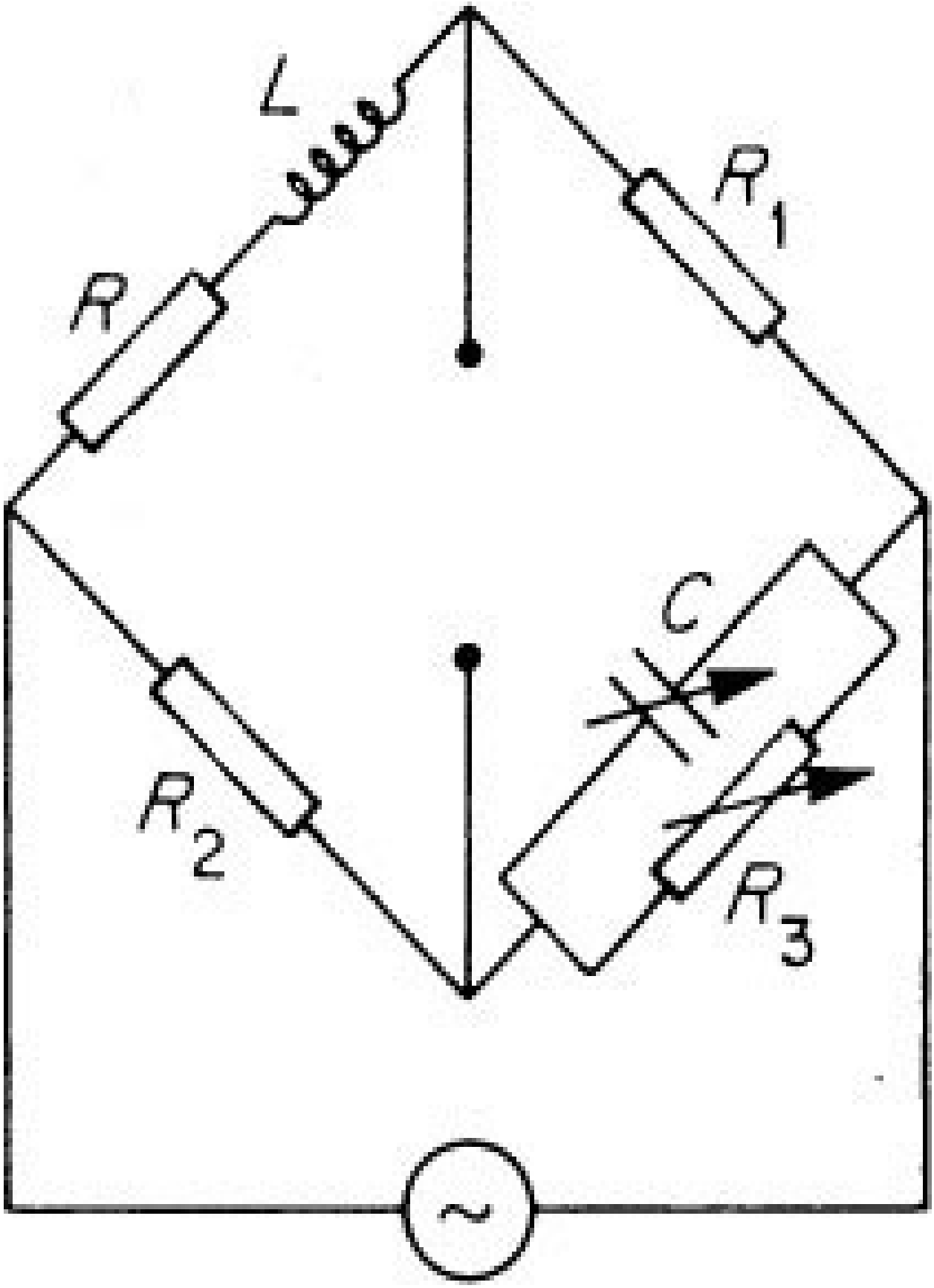
8.4 The bridge in Figure 8.28 is a phase-shifting network. Show that if the generator voltage V has a constant amplitude, the voltage across XY has constant amplitude but varies in phase as R alters. Sketch the phase variation as R changes from 0 to ∞.

Figure 8.28.


R_0
R_0
X
Y
R
C
V
~

8.5 Derive the balance conditions for the Maxwell inductance-capacitance bridge in Figure 8.29.

Figure 8.29.


L
R
R1
C
R2
R3
~

8.6 At resonance the impedance of an LCR circuit is real, and a balance can be found in a resonance bridge containing only resistors in three of its arms. What are the balance conditions for the bridge in Figure 8.30?

8.7 A mutual inductance M couples two ideal self-inductances L as shown in Figure 8.31. Set up the loop equations for this circuit, including the e.m.f. induced in each loop caused by the alternating current in the other. Solve to find the resonant angular frequencies of the circuit, i.e. the angular frequencies at which the impedance seen by the generator is zero.

Figure 8.30.

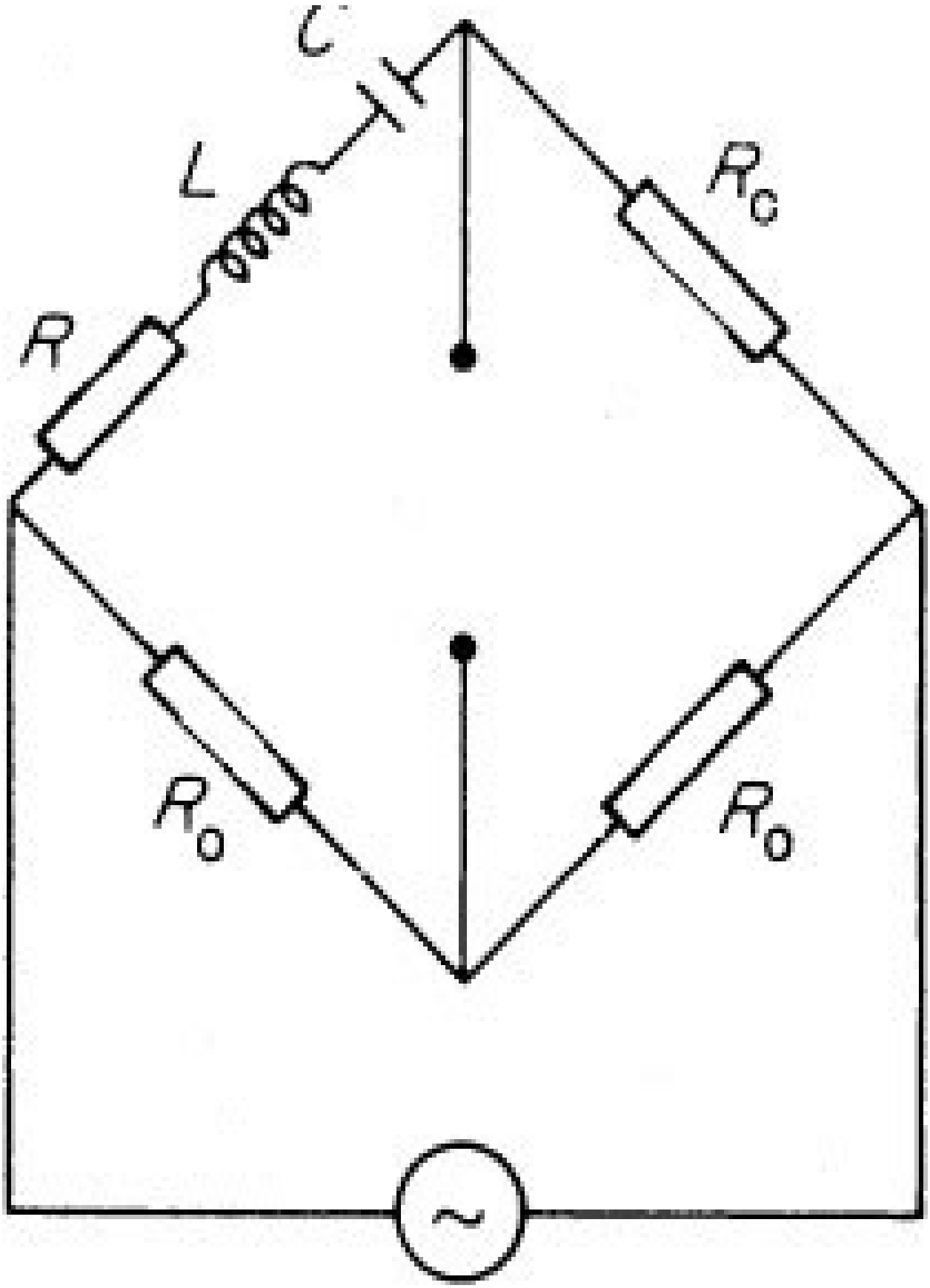
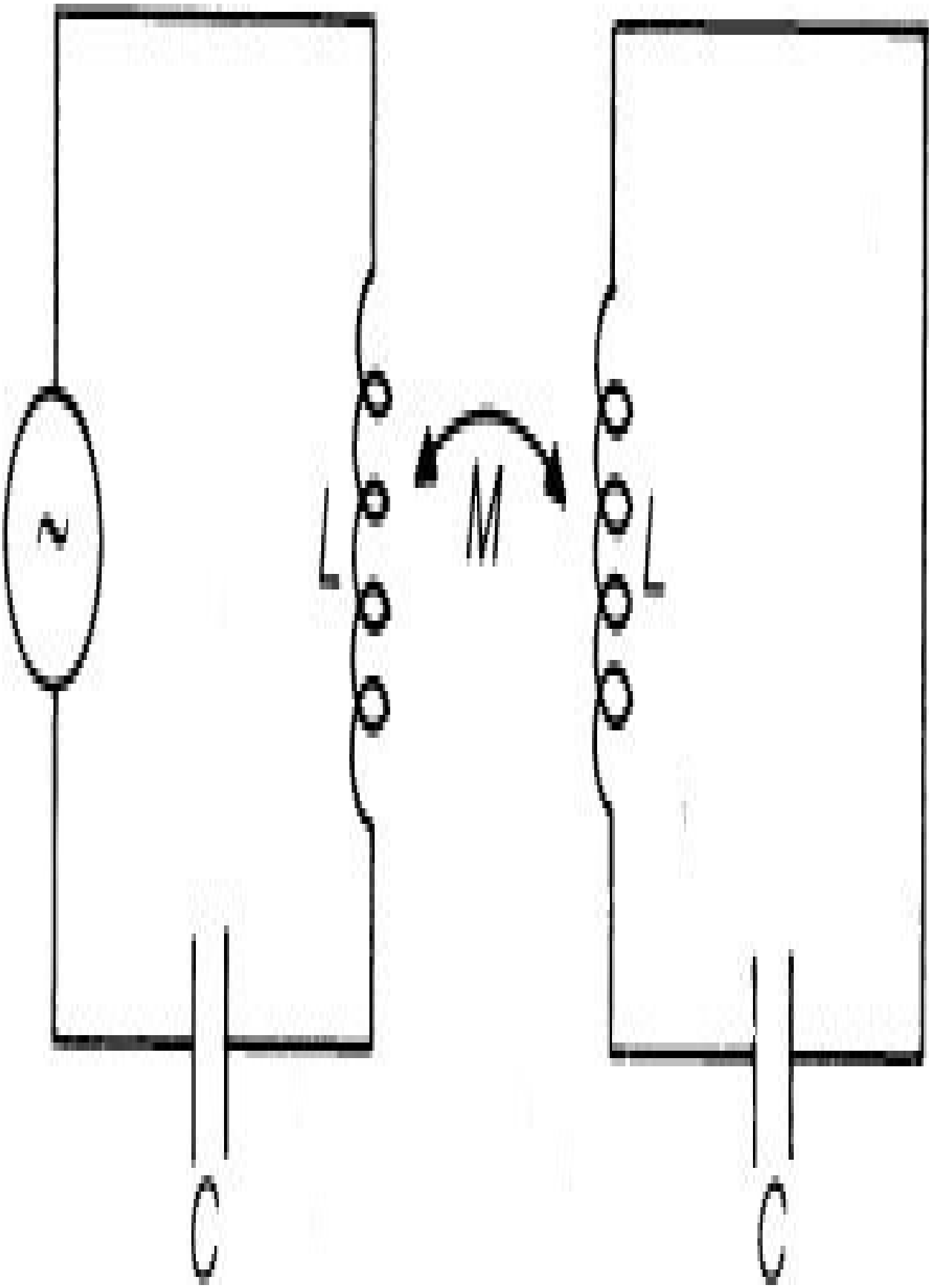

C
L
R
R_C
R_0
R_0
~

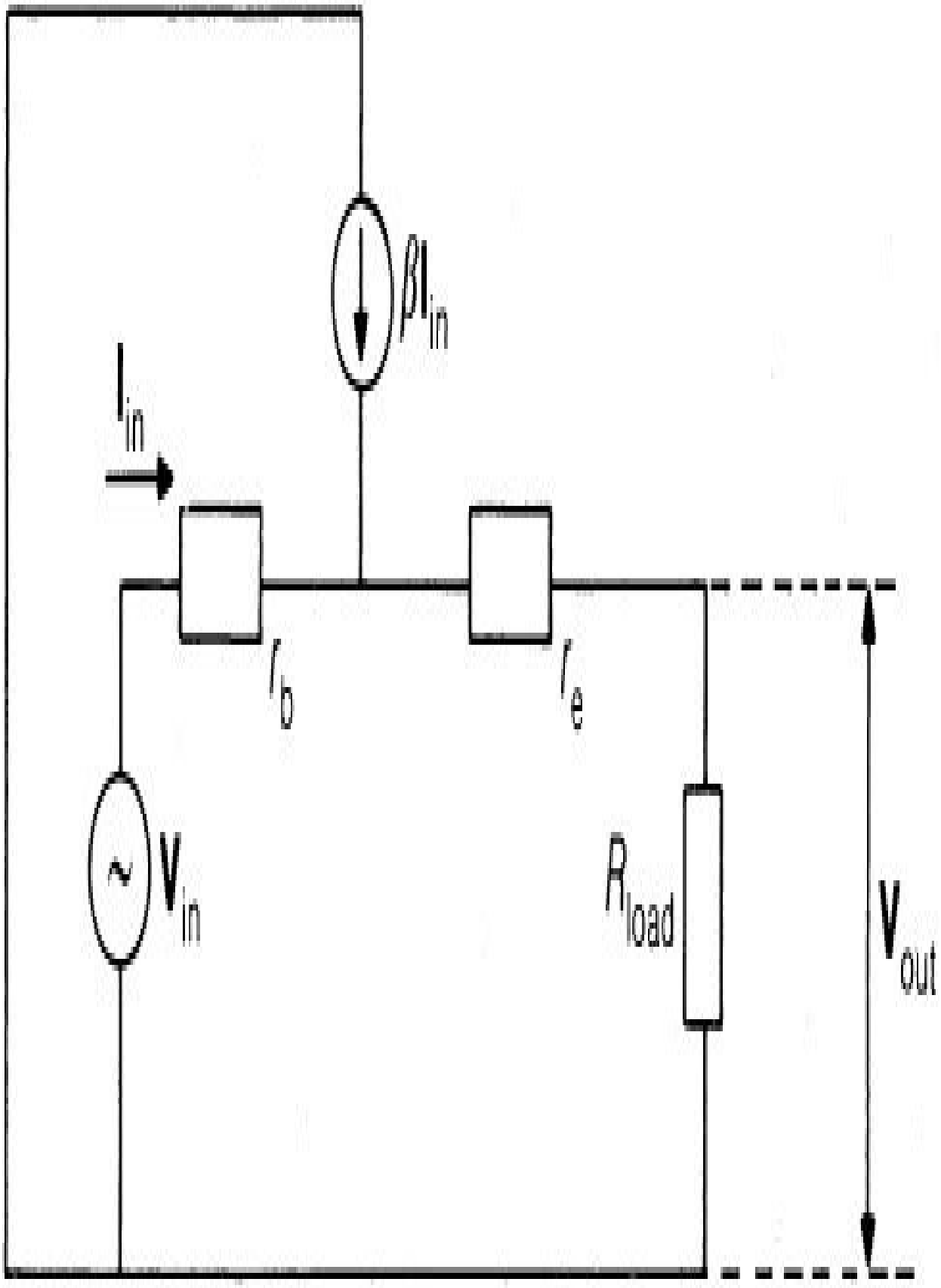
Figure 8.31.


~
L
M
L
C
C

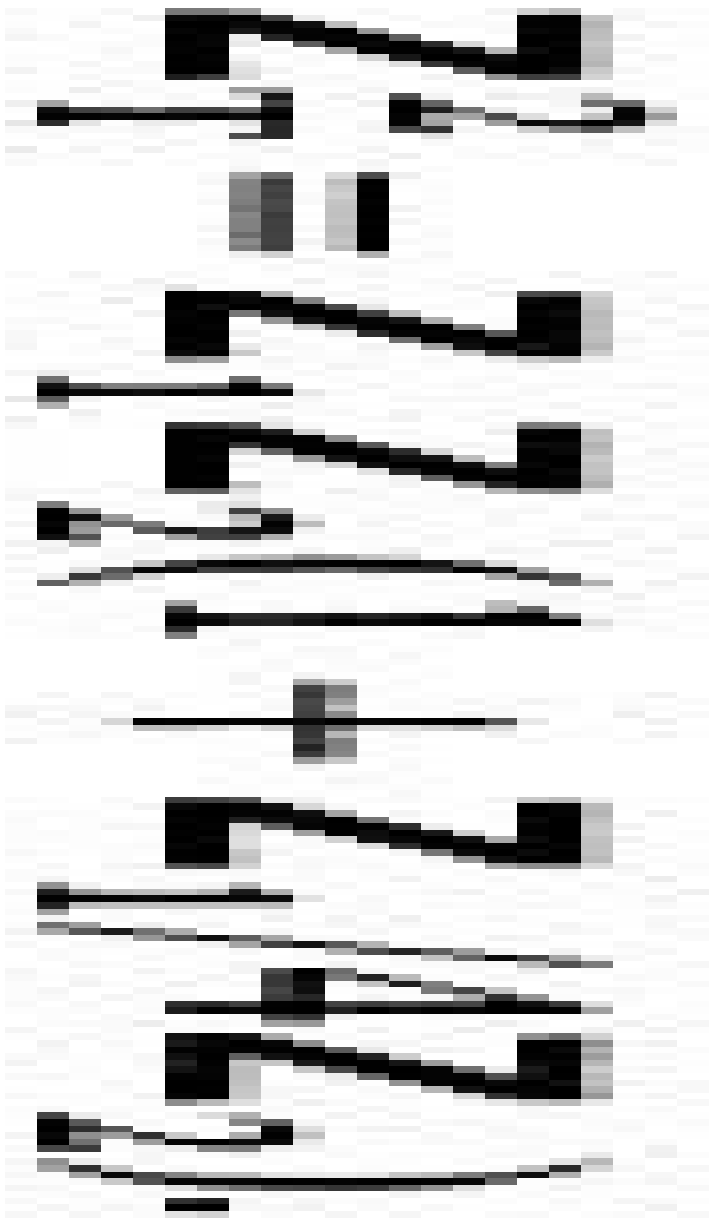
8.8 An A.C. generator of voltage V and internal impedance (X + jY) drives current through a load of impedance Z. Show that maximum power transfer to the load occurs when Z = X – jY. (Hint: the power is dissipated by the real part of the load impedance. For a fixed value of the real part Z, choose the imaginary part to maximize the current amplitude, then vary the real part for maximum power transfer.)

8.9 The network in Figure 8.32 is an equivalent circuit for the transistor circuit known as the emitter follower. The transistor acts as a current amplifier with gain $\beta \gg 1$, driving a current $(\beta + 1)\, I_{in}$ through the resistance r_e and the load resistor R_{load}. Find the open-circuit output voltage and the short-circuit output current of the emitter follower, and hence its output impedance.

Figure 8.32. An equivalent circuit for the emitter follower.


βI_in
I_in
r_b
r_e
~
V_in
R_load
V_out

8.10 The ladder networks described in section 8.4.1 may be split up into symmetrical ‘π-sections’ or ‘T-sections’ as indicated in Figure 8.33. Show that the matching impedances of the π- and T-sections are given by



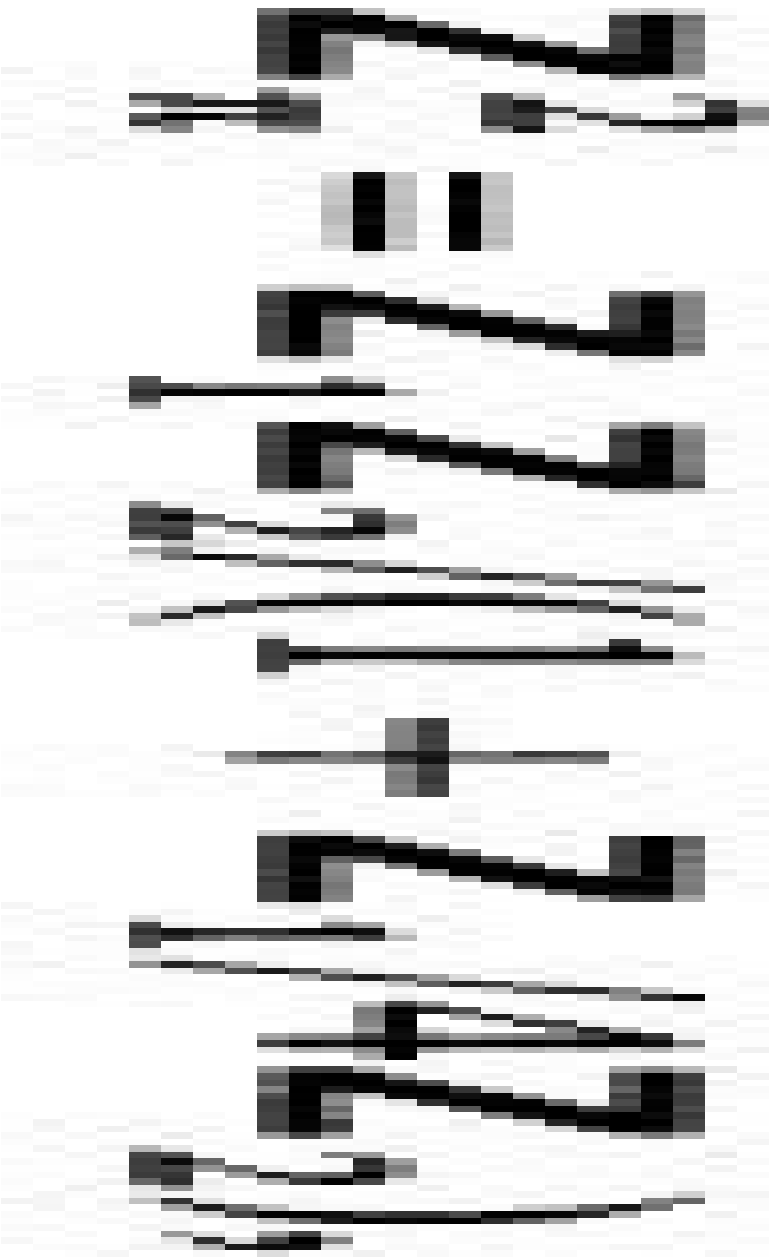
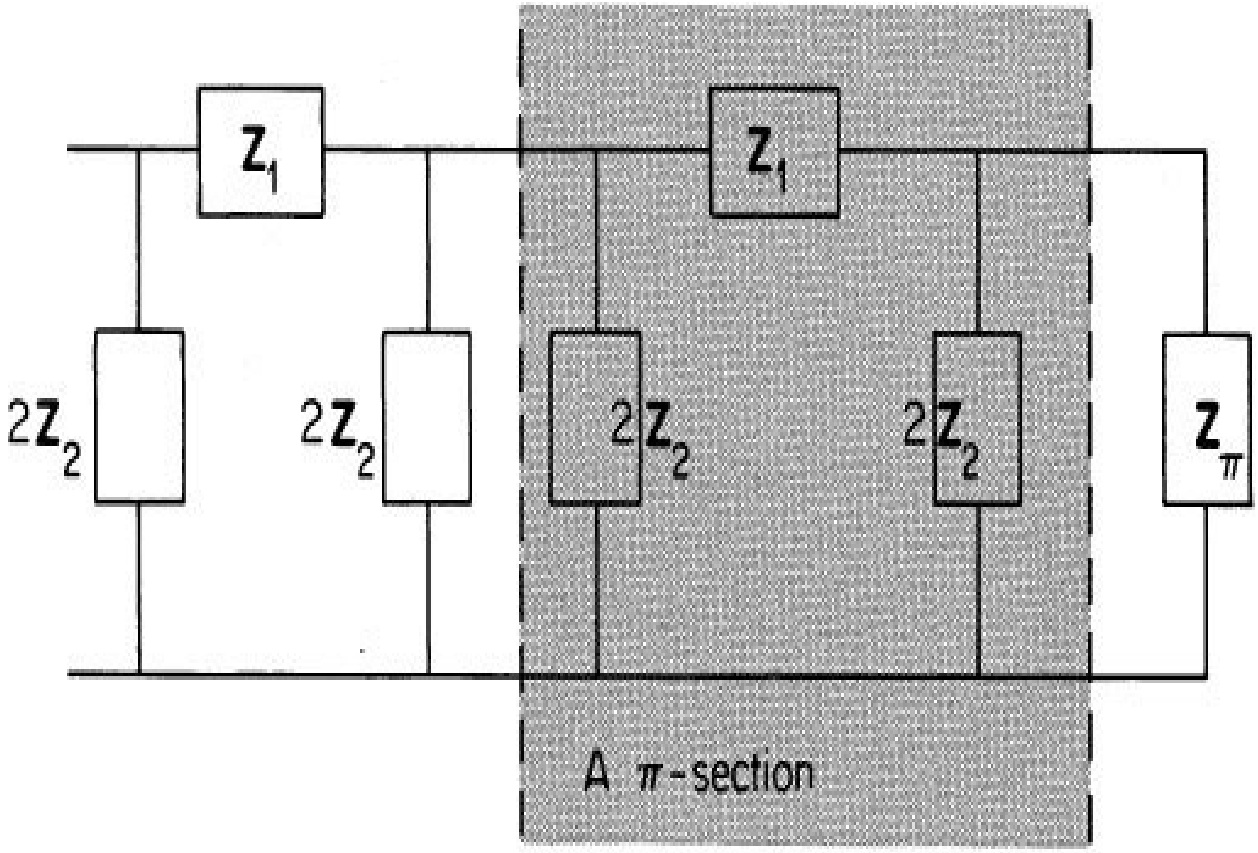
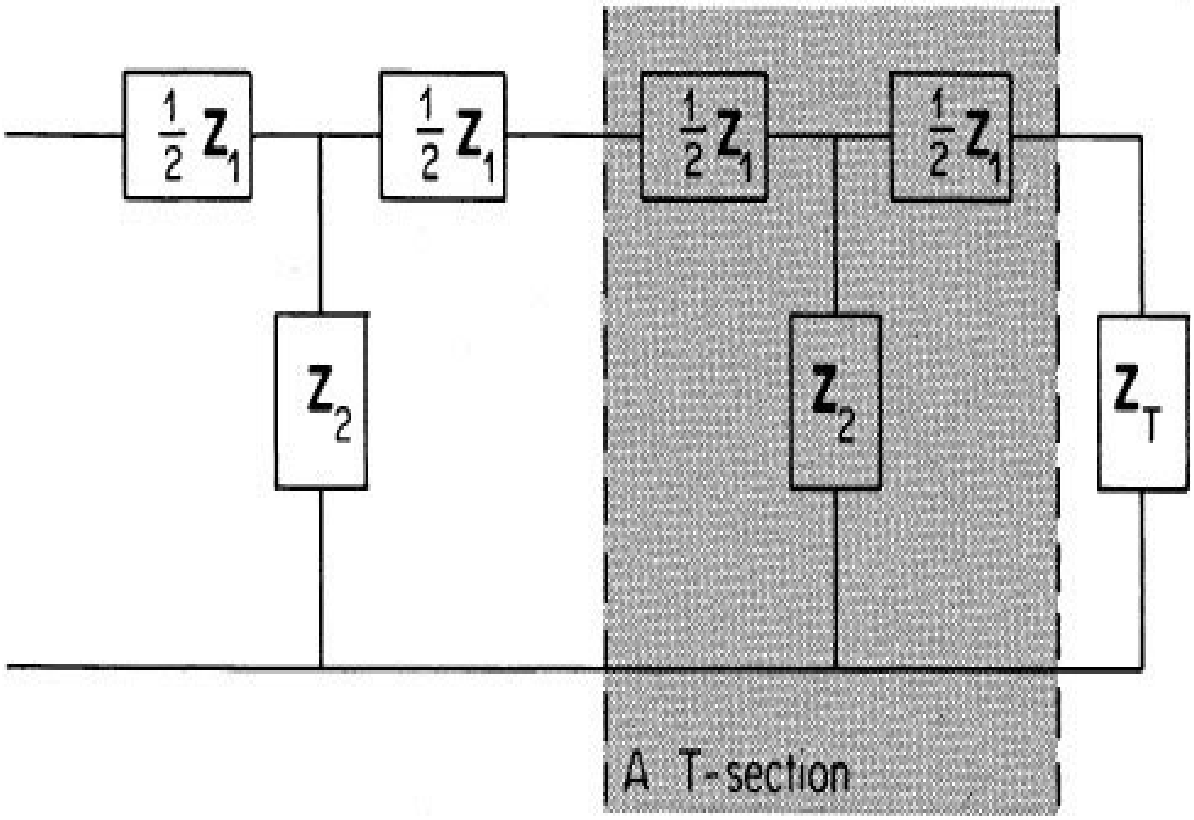


Figure 8.33.


Z_1
Z_1
$2Z_2$
$2Z_2$
$2Z_2$
$2Z_2$
Z_π
A π-section


$\frac{1}{2}Z_1$
$\frac{1}{2}Z_1$
$\frac{1}{2}Z_1$
$\frac{1}{2}Z_1$
Z_2
Z_2
Z_T
A T-section

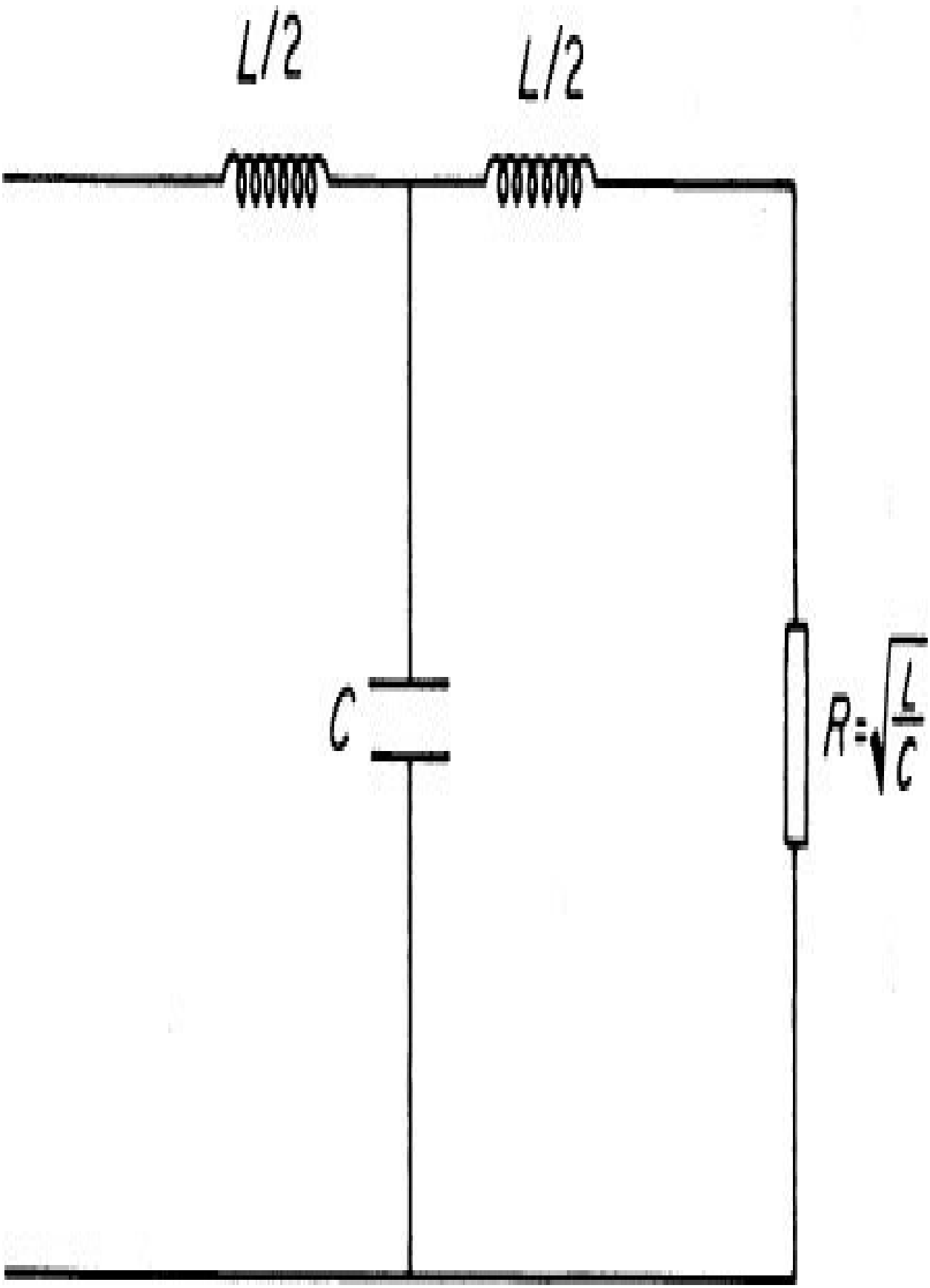
8.11 Calculate the voltage transfer function for the T-section network in Figure 8.34.

8.12 An A.C. generator with output impedance 16Ω is connected to the primary of an ideal transformer. The secondary drives current through a 100Ω load. What should be the turns ratio of the transformer to achieve the maximum power dissipation in the load?

8.13 The primary of a transformer presents an inductance of 1H when the secondary is open-circuited. If a generator is connected to the primary and a 1000 pF capacitor across the secondary, the current drawn from the generator has a resonant maximum at.a frequency $(10^4/2\pi)$Hz. The secondary has ten times as many turns as the primary. What is the coupling coefficient of the transformer coils?

8.14 Use Faraday's law to show that the eddy current density in one of the soft iron sheets making up a laminated transformer core is proportional to the distance from the central plane of the sheet. Hence, assuming that the magnetic field $B = B_m \cos \omega t$, derive the expression $\sigma \omega^2 a^2/24$ for the power density in sheets of thickness a and conductivity σ. If the 50 Hz mains transformer in Figure 8.24 has laminations 1 mm thick with conductivity 2.5×10^6 $(\Omega m)^{-1}$, what is the eddy current power loss when $B_m = 1T$?

Figure 8.34.


L/2
L/2
C
$R=\sqrt{\frac{L}{C}}$

* Notice that this transfer function has sharp peaks, which must lead to distortion of non-sinusoidal signals. Somewhat better performance may be achieved with more complex networks than the equal-section filter considered here. However, filters made up of passive elements (inductors, capacitors and resistors) always have undesirable peaks in their transfer functions. Such peaks may only be eliminated by the use of active filters, which incorporate amplification and feedback.

CHAPTER 9

Transmission lines

In discussing A.C. circuits, we have assumed previously that current and voltage changes are transmitted instantaneously along the leads connecting different components in a network. At a given moment, a single value has been assigned to the current throughout the length of any branch of the network. This cannot be exactly true, since we know that no signal can travel faster than the speed of light. What has not been taken into account is the stray impedance associated with the leads (i.e. the self-inductance of the leads and their mutual inductance and capacitance with other parts of the circuit). When the stray impedances are included, they introduce phase changes and time delays, just as coils and capacitors introduce phase changes and time delays between the sections of the ladder networks described in section 8.4.1.

The value of stray impedances depends on the shape of each particular circuit, but their effect is always such that voltage and current changes move about the circuit at a speed not more than the speed of light. The time delays caused by strays are negligible at low frequencies, because they are then much shorter than the period of the alternating voltages and currents. At what frequency can the strays no longer be ignored? This question has already been raised in the discussion of networks in section 8.1.3. It was pointed out there that in ordinary laboratory circuits, network analysis breaks down at a frequency of about 1000 MHz. At such a high frequency, the magnitude of the impedances of stray capacitances and stray inductances are similar. All components are therefore either shunted by current carried by stray capacitance, or cause voltage drops no bigger than the voltage drops due to the stray inductance of leads. Although this argument may not seem to have anything to do with the speed of light, it is roughly equivalent to saying that at 1000 MHz the period of alternating voltages is comparable to the time taken for a light signal to cross the circuit. In a time of

10^{-9} sec, equal to one complete period at 1000 MHz, light travels 30 cm, about the size of the circuits we had in mind.

It must be emphasized that there is no well-defined frequency limit below which lead impedances may be neglected. If a signal is to be carried over a distance of a hundred metres, for example, then propagation delays are important at only 1 MHz. High frequency signals can be carried over such distances along a transmission line. A transmission line consists of a pair of parallel conductors which have the same cross-section anywhere along the length of the line. Between the conductors there is a capacitance which, apart from end effects, is proportional to the length of the line. A transmission line with a particular cross-section therefore has a definite capacitance per unit length between its conductors. Similarly, if the conductors are joined together at both ends of the line, a closed circuit is formed which has a self-inductance proportional to the length of the line. The line thus has a definite self-inductance per unit length. Because of the time taken to accumulate charge on the capacitance and to build up current across the inductance, delays are introduced, and signals propagate along the line at a constant speed: it turns out that this speed equals the speed of light in the medium in which the line is embedded.

As they travel along the line, signals are attenuated because of energy losses in the conductor (which has a non-zero resistance) and in any dielectric materials which may be present. In practice the attenuation is usually small and it is often a very good approximation to assume that a transmission line is lossless, so that it causes no attenuation at all. The properties of a lossless line, which are entirely determined by its capacitance per unit length and self-inductance per unit length, are first discussed in a general way in section 9.1. Then in section 9.2 some commonly used types of line are described. When a line is used to join two different circuits together, signals may sometimes be transmitted unaltered from one circuit to the other, but in general they will be partially reflected at the ends of the line. The amount of reflection depends on the values of the impedances terminating the line: this is explained in section 9.3, and the input impedance of a mismatched line is derived in section 9.4. Finally, the attenuation caused by resistive and dielectric losses is evaluated in section 9.5.

9.1 PROPAGATION OF SIGNALS IN A LOSSLESS TRANSMISSION LINE

A lossless line is made up of conductors which have zero resistance, and which are separated by a perfect insulator. No energy is dissipated in the line, whatever the frequency of the voltage across the conductors. Suppose that a lossless line has a capacitance per unit length C and a self-inductance per unit length L. Let us terminate a section of the line by joining an impedance Z across the conductors at one end. Viewed from the other end of the line, the circuit formed by the line plus its terminating impedance Z has in input impedance Zin, say. If an extra length δx is added to the line, the network has an additional self-inductance $L\delta x$, and there is an additional capacitance $C\delta x$ shunting the conductors, as shown in the equivalent circuit in Figure 9.1. This figure is reminiscent of the ladder networks discussed in the previous chapter. A transmission line is a limiting case of a low-pass filter network in which the inductance and capacitance of each section have been made infinitesimally small; the cut-off frequency is therefore indefinitely large, that is to say, the line will transmit signals of any frequency without attenuation. At low frequencies the low-pass filter has a characteristic impedance Z0 = (see section 8.4.1), and each section introduces a time delay (see Equation (8.23)). By analogy we should expect the line to have a characteristic impedance Z0 = at all frequencies, so that in Figure 9.1, if the line is terminated by an impedance Z0, then for all frequencies Zin = Z0, no matter what is the length of the line. Also the time delay introduced by an extra length δx of the line should be δx , at any frequency, i.e. signals are propagated along the line at a speed 1/ . These values are correct, as we shall now verify by setting up the differential equations for the voltage and current on the line.

Consider the length δx of the line shown in Figure 9.2. At a time t the voltage and current at the position x on the line are V(x, t) and I(x, t) respectively. Over the distance δx the voltage changes to $V(x + \delta x, t)$ because there is a voltage drop across the self-inductance $L\delta x$ of the intervening section of the line, and we have

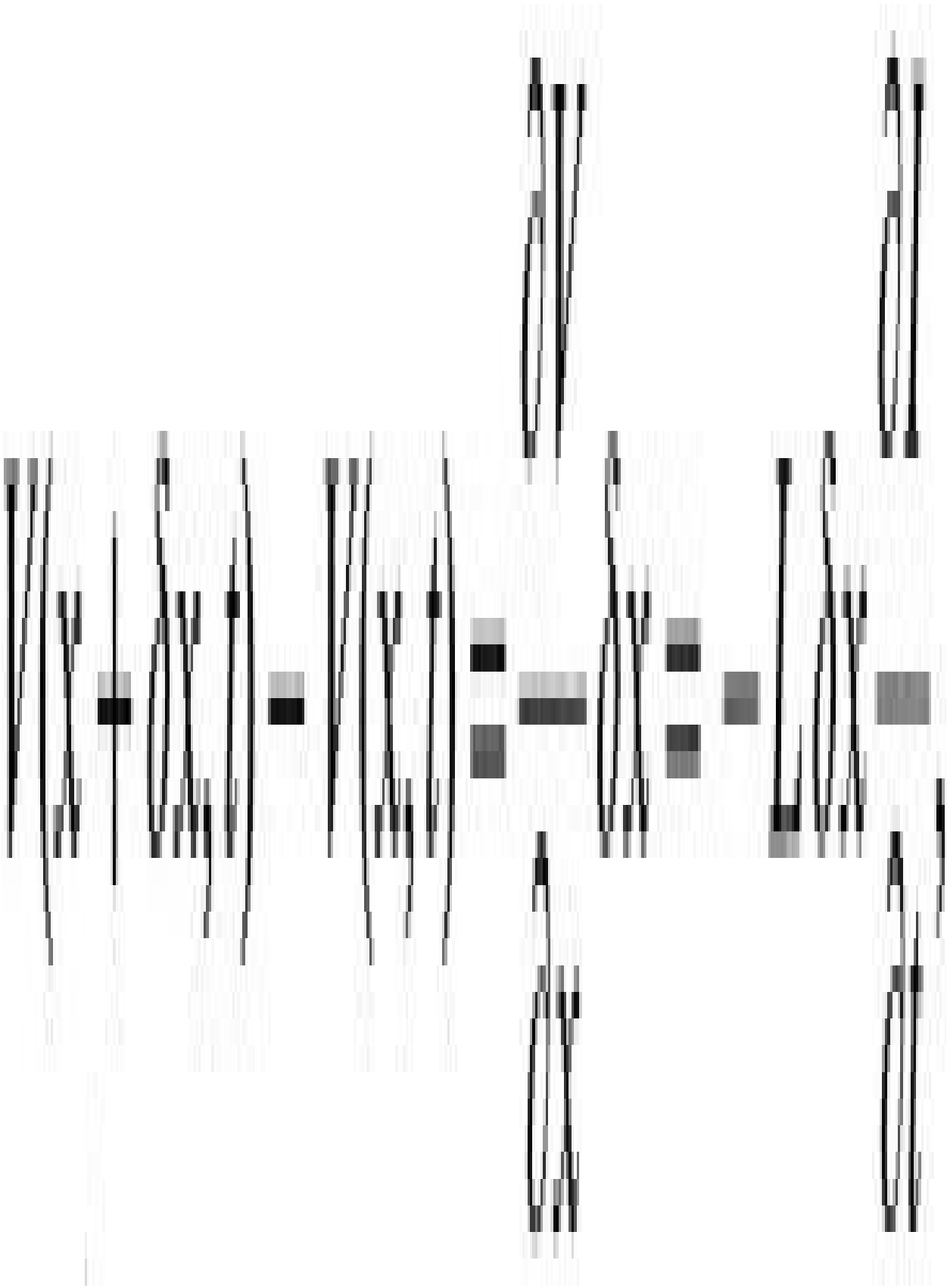


Figure 9.1. The equivalent circuit for a length δx, of transmission line.

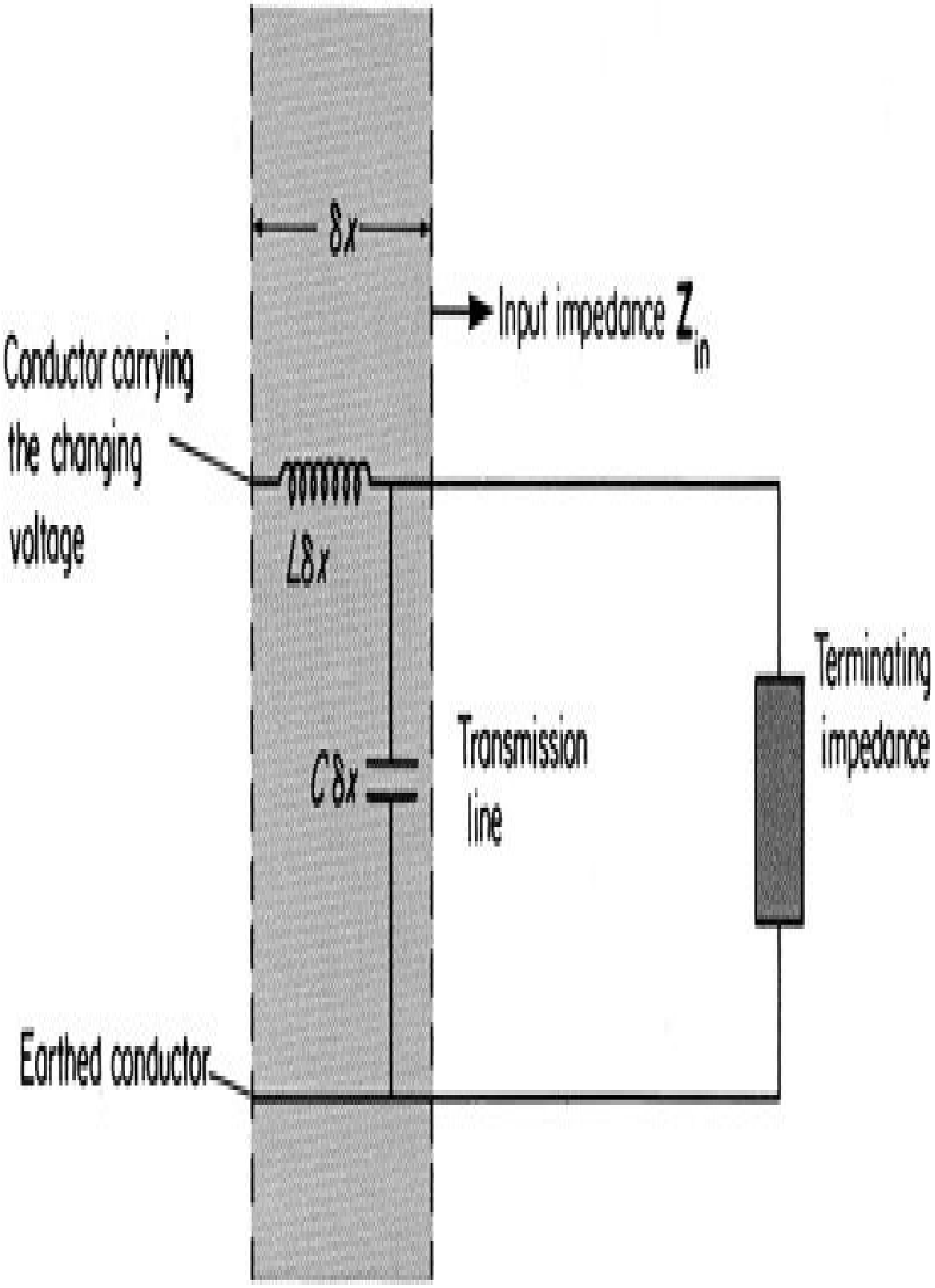
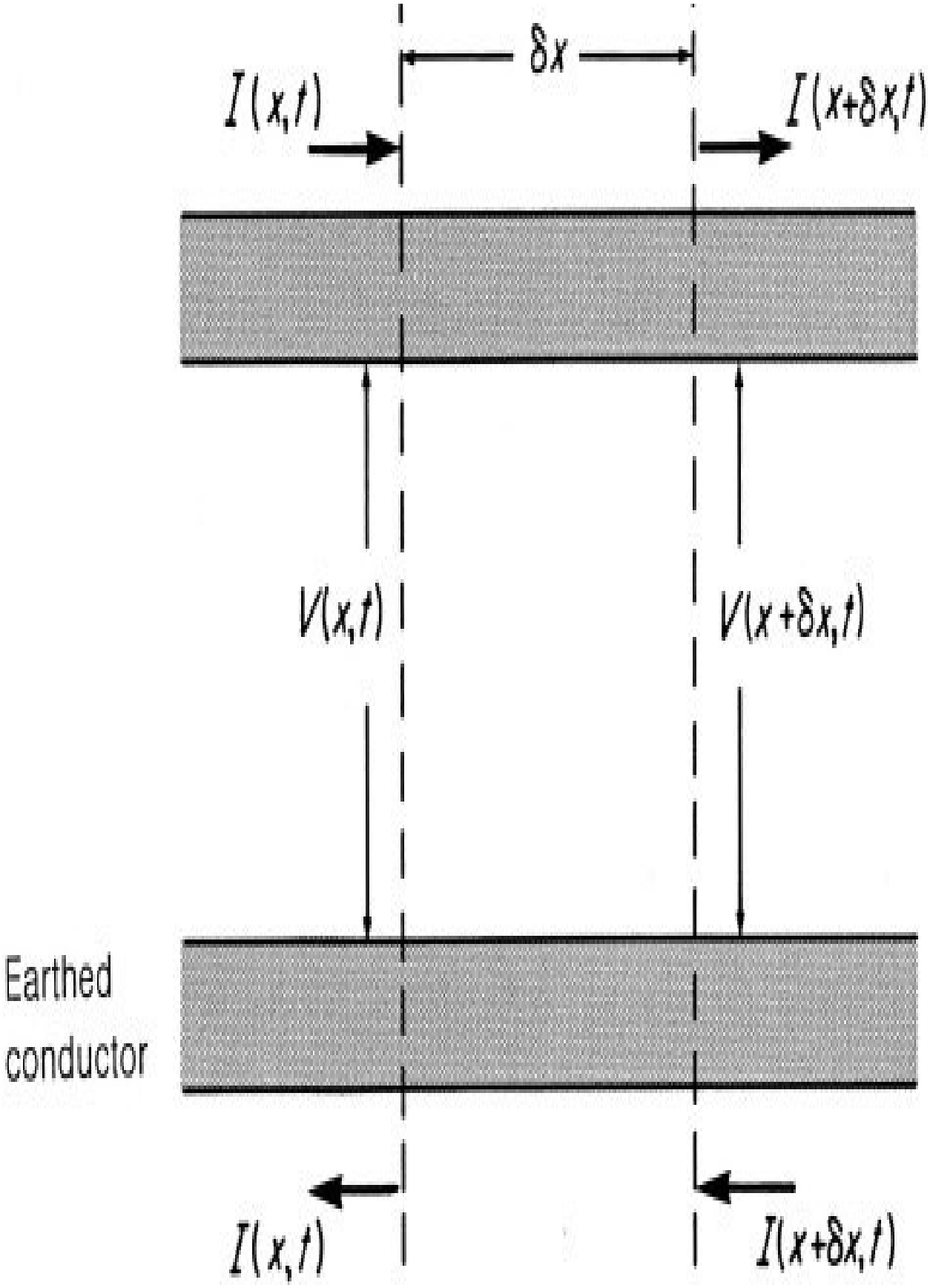

δx
Input impedance Z_{in}
Conductor carrying the changing voltage
Lδx
Cδx
Transmission line
Terminating impedance
Earthed conductor

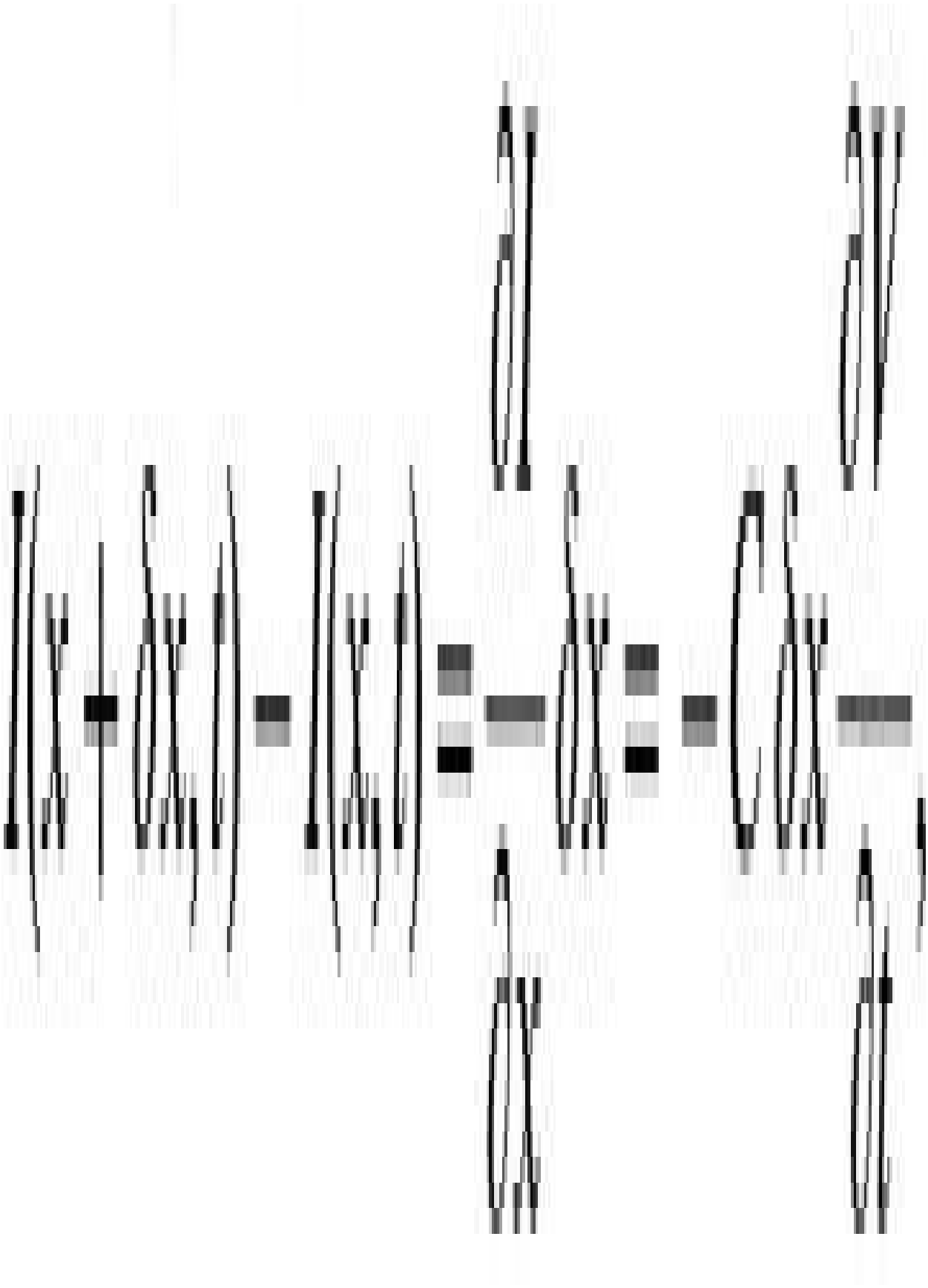
Figure 9.2. The voltage and current on a transmission line are functions of position and of time.


δx
I(x,t)
I(x+δx,t)
V(x,t)
V(x+δx,t)
Earthed
conductor
I(x,t)
I(x+δx,t)

or

(9.1)

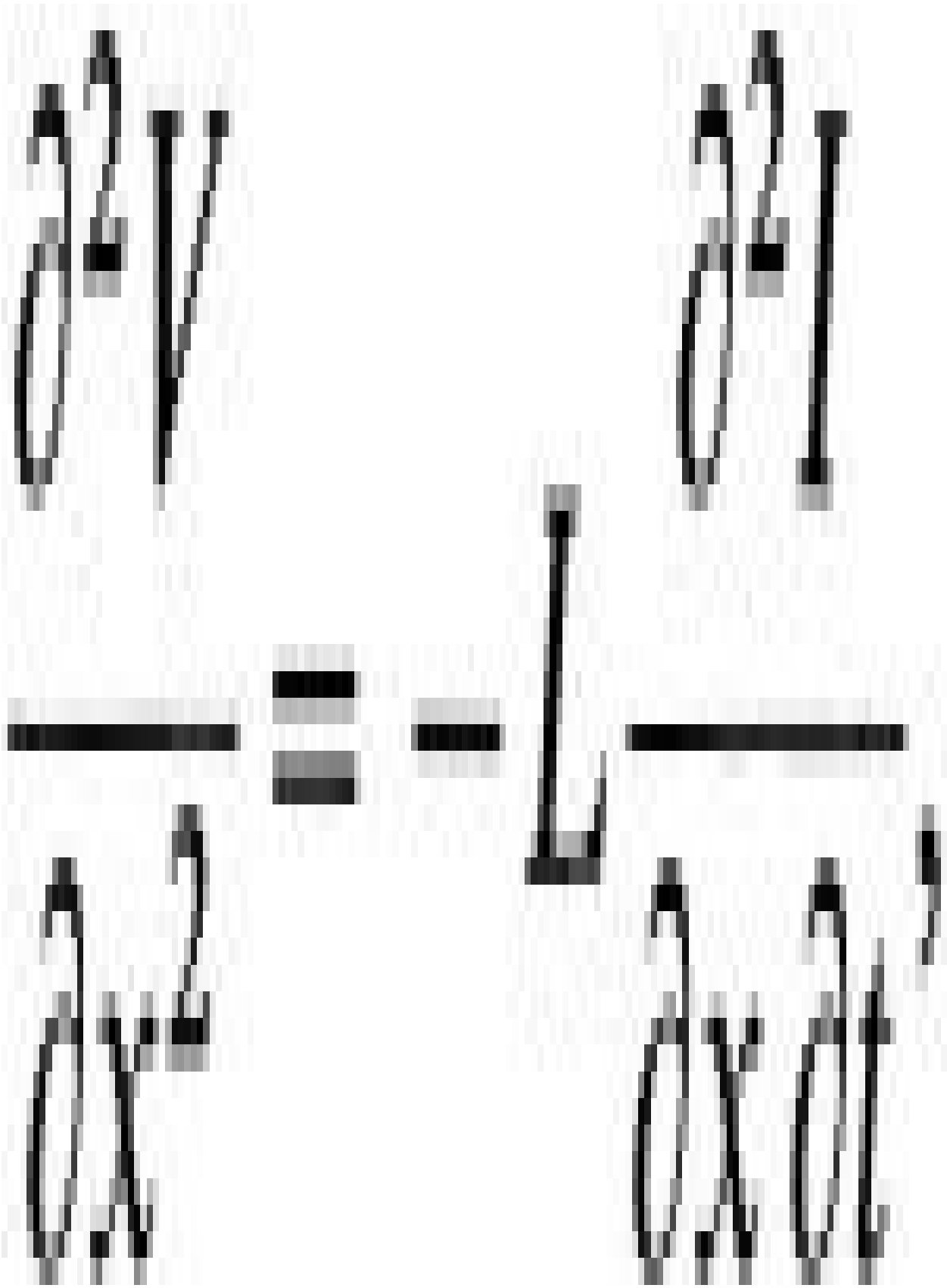
The current flowing in the two conductors also varies with x, because of the capacitance between them. (Notice that the symmetry of the problem implies that in a lossless line equal and opposite currents are flowing in the conductors at any x.) The capacitance $C\delta x$ in the length δx carries a current $C\delta x(\partial V/\partial t)$, and hence



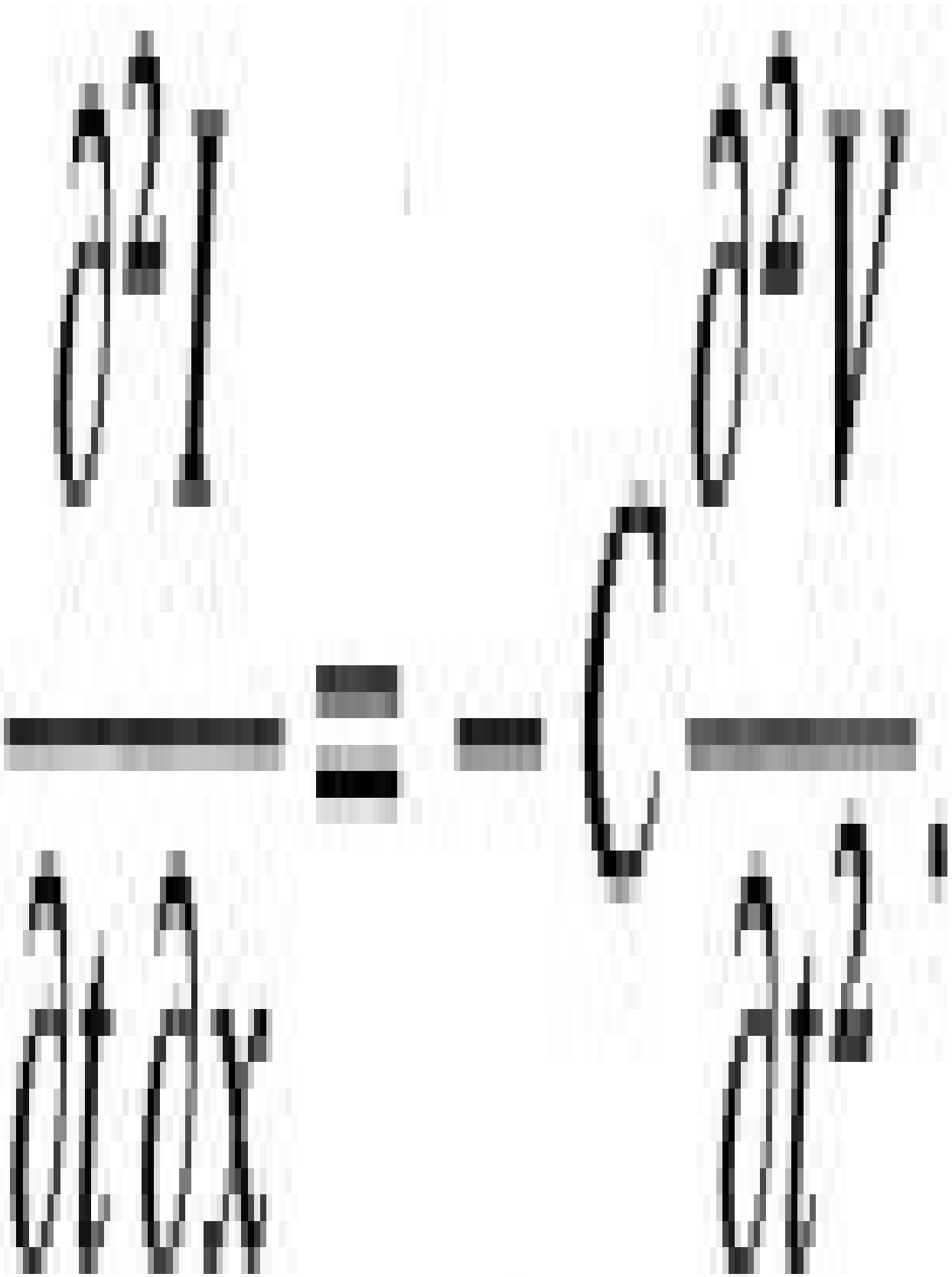
or

(9.2)

Differentiating Equation (9.1) with respect to x and Equation (9.2) with respect to t,



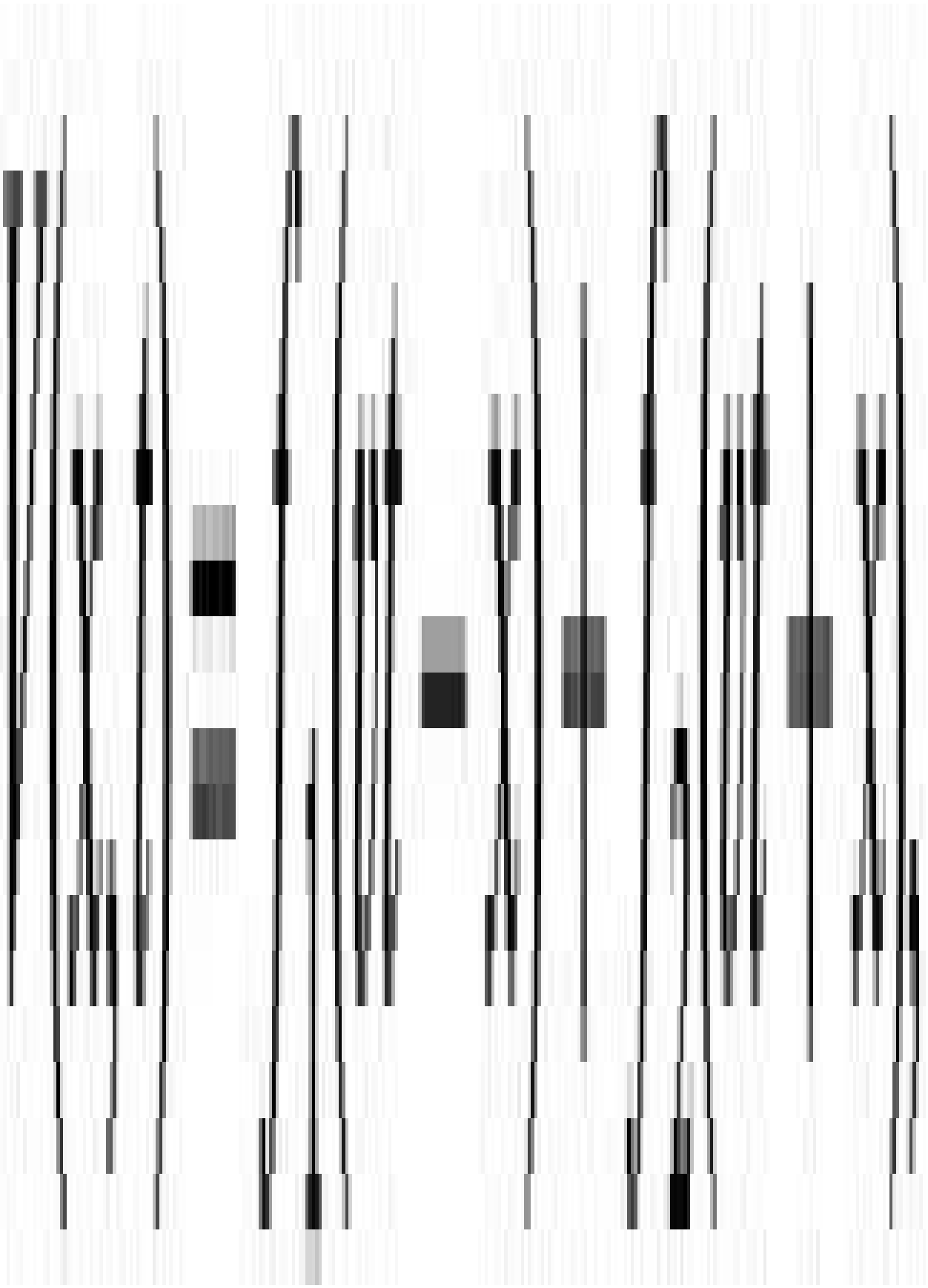
and



Since x and t are independent variables, the order of differentiation is immaterial, and the two equations above reduce to

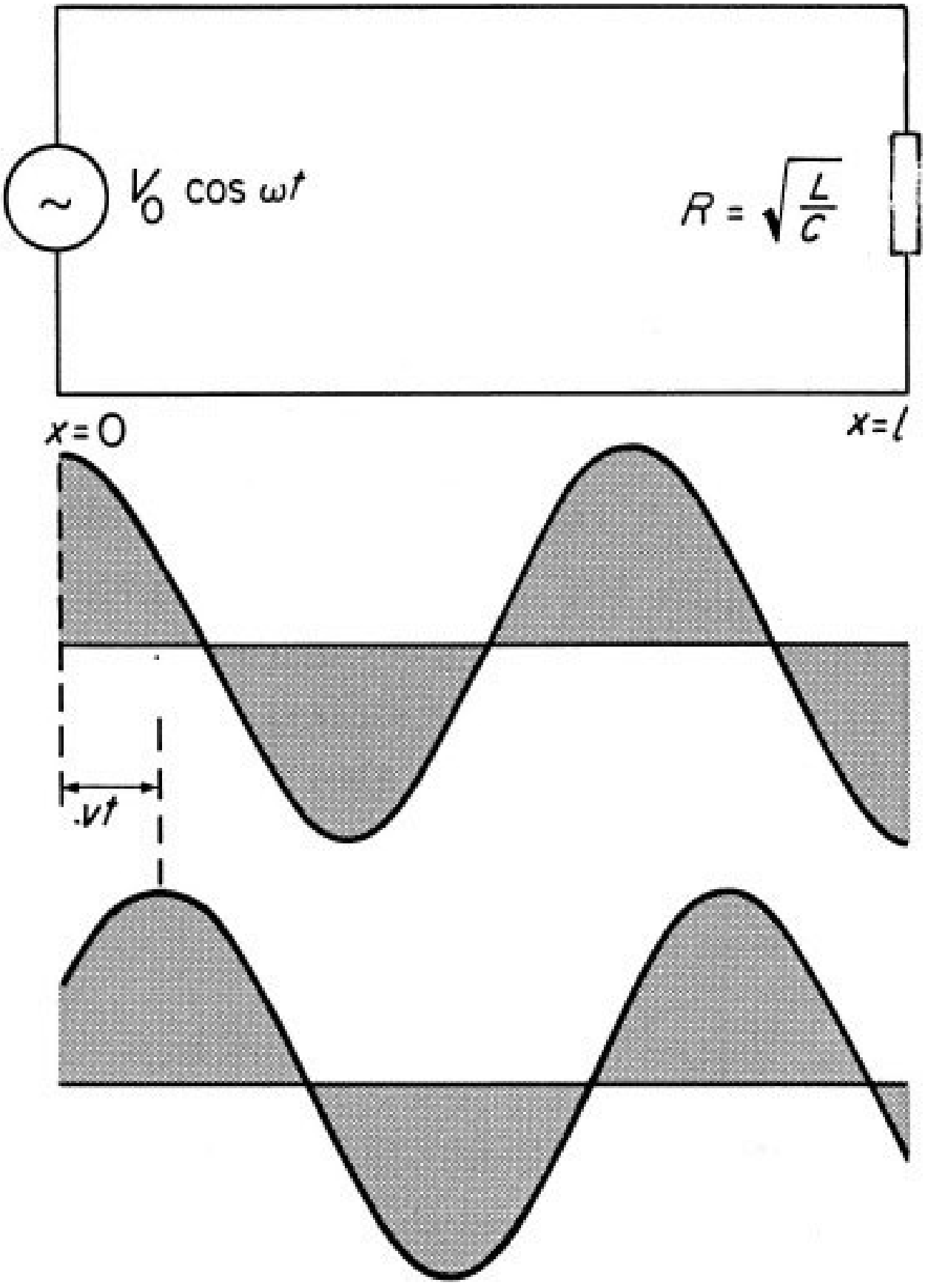
(9.3)

This equation has the same form as the differential equation describing the transverse displacement of a stretched string. Just as waves travel along a string when one end is vibrated, so voltage waves travel along the transmission line when a changing voltage is applied to it. Equation 9.3 is a wave equation in one dimension. The most general way of writing its solution is



where f1 and f2 are any functions whatever, and v = 1/ . This can be verified by substituting for V in Equation (9.3). The function f1 represents a wave travelling along the line from left to right at a speed v in the positive x-direction, and f2 represents a wave travelling from right to left. For example, if V(x, t) = V0cos(vt − x), then at t = 0 the voltage on the line has the sinusoidal variation shown in Figure 9.3. Later on, at time t, the voltage pattern has the same shape, but it has been shifted to the right through the distance vt.

Figure 9.3. A wave travelling along a matched length of transmission line satisfies Ohm’s law for the resistor at the end of the line at all times.


~
$V_0 \cos \omega t$
$R = \sqrt{\frac{L}{C}}$
$x = 0$
$x = l$
$v t$

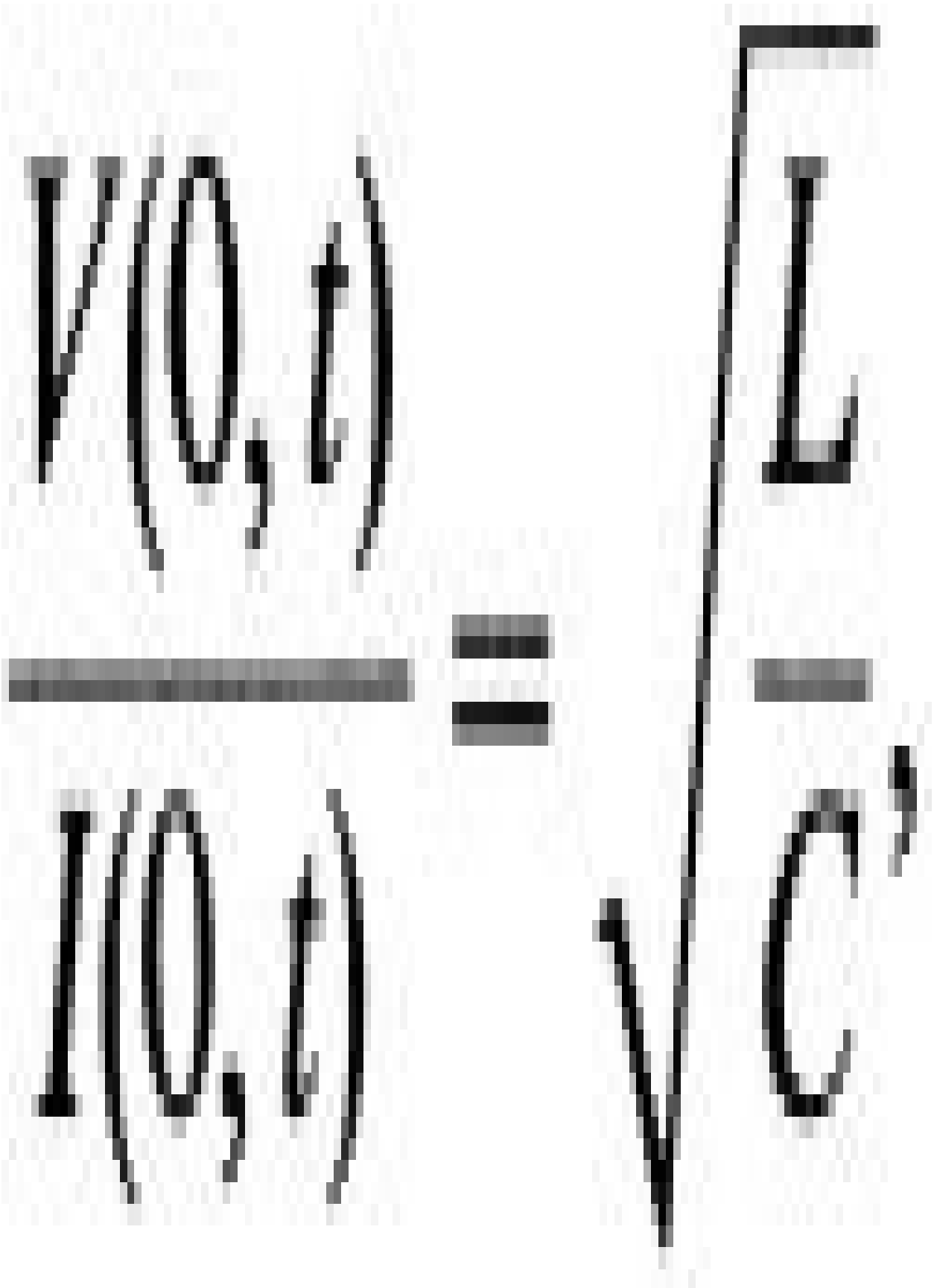
Instead of eliminating the current from Equations (9.1) and (9.2), the voltage V can be eliminated, leading to a wave equation for the current:

(9.4)

This current must also be a superposition of waves travelling with speed 1/ in either direction along the line. For the moment we will restrict attention to waves of a single angular frequency ω, travelling from left to right. Consider the line of length l shown in Figure 9.3. At the end where x = 0 a generator maintains an alternating voltage of angular frequency ω and amplitude V0. The other end of the line at x = l is terminated by a resistance . Waves travel away from the generator, from left to right, and the voltage and current on the line are

(9.5)

Here the wave-number k = ω , since the solution must represent waves travelling away from the generator at a speed 1/ . This solution satisfies both the line Equations (9.1) and (9.2), and also satisfies Ohm's law V = I/ at the end of the line x = l terminated by the resistor. The input impedance of the line is



whatever the length of the line, i.e. the characteristic impedance of the line is . Notice that although the line has inductance and capacitance, its characteristic impedance is purely resistive, so that current and voltage are in phase with each other when the line is matched.

Waves travelling from right to left are also solutions of the wave Equation (9.4). These have the form

(9.6)

The minus sign in the expression for the current arises because the direction of the current on the unearthed conductor, as defined in Figure 9.2, is in the opposite direction to the wave. The voltage and current in Equation (9.6) also satisfy Equations (9.1) and (9.2) everywhere on the line. Clearly, if the line were terminated at its input end by its characteristic impedance, the impedance looking at the line from its right-hand end is also . In other words, the output impedance of a transmission line matched at the input is .

The above argument holds for any angular frequency. We have shown that any alternating voltage is propagated along the line at a speed 1/ , and that the characteristic impedance of the line is real and equal to at all frequencies. Since the amplitudes of the voltage and current in Equation (9.5) are independent of x, the lossless line transmits all signals without attenuation and without distortion.

9.2 PRACTICAL TYPES OF TRANSMISSION LINE

In this section we calculate the distributed parameters—the capacitance per unit length and the inductance per unit length—of some commonly used transmission lines. From these parameters are derived the characteristic impedances of the lines and the speeds at which signals are propagated along them.

9.2.1 The parallel wire transmission line

Figure 9.4 is a cross-section of a transmission line made of a pair of similar cylindrical wires, each of radius a and with their centres separated by a distance 2d. The wires are embedded in a medium of relative permittivity ε and relative permeability μ. The capacitance per unit length of a pair of wires has already been calculated in section 3.3.1. If 2d » a, it is approximately

(9.7)

To find the inductance per unit length, we make a closed circuit by joining the ends of a section of length l. The self inductance of this whole circuit is the magnetic flux through the circuit when it is carrying a current of 1 amp. The field outside a long straight wire carrying I amps, at a distance r from its centre, is μμ0I/(2πr) (Equation (4.37)). If the direction of the current is as shown in Figure 9.5 both wires contribute a field B pointing into the figure, and the total

Figure 9.4. Cross-section of a parallel wire transmission line.

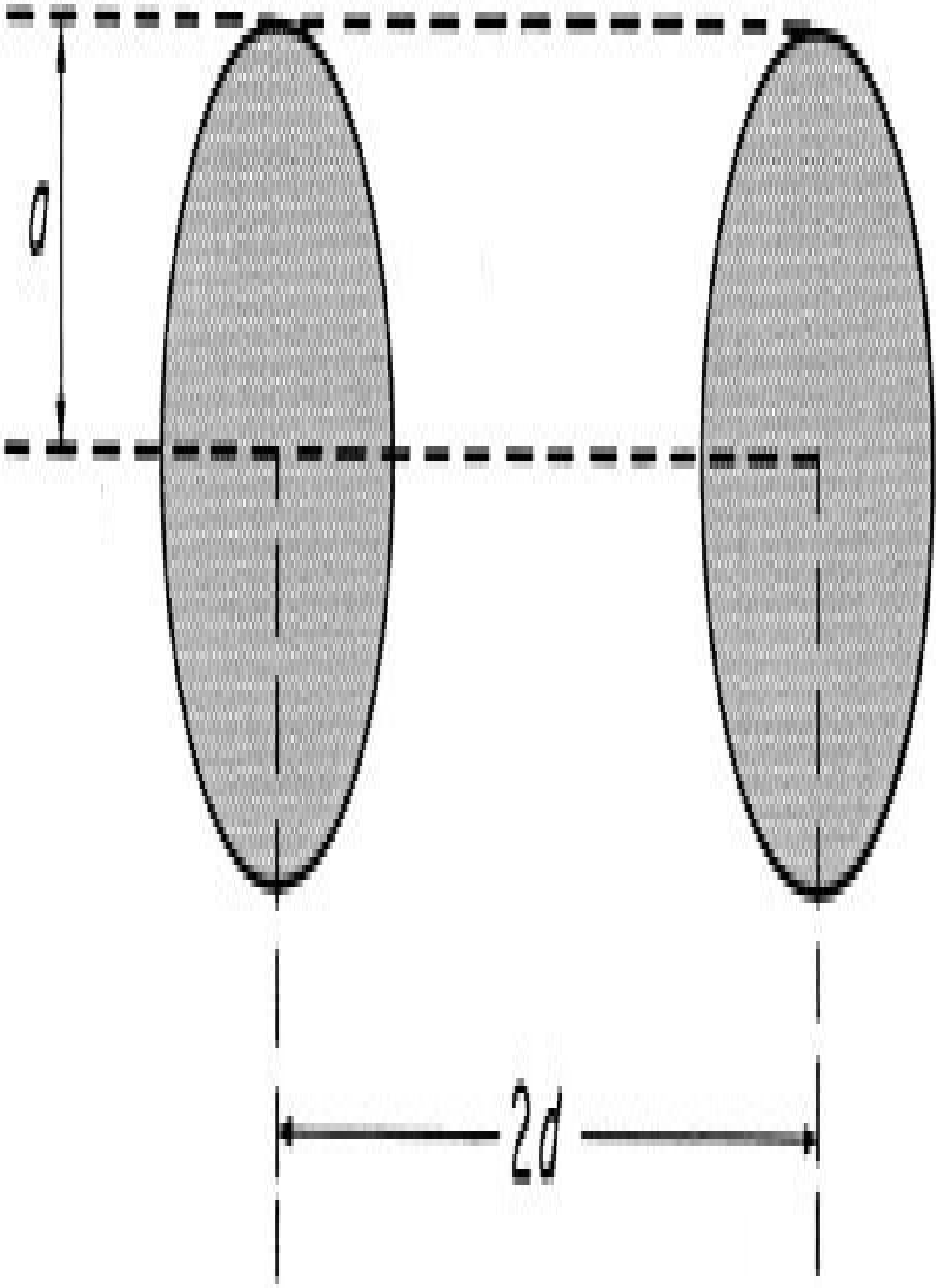
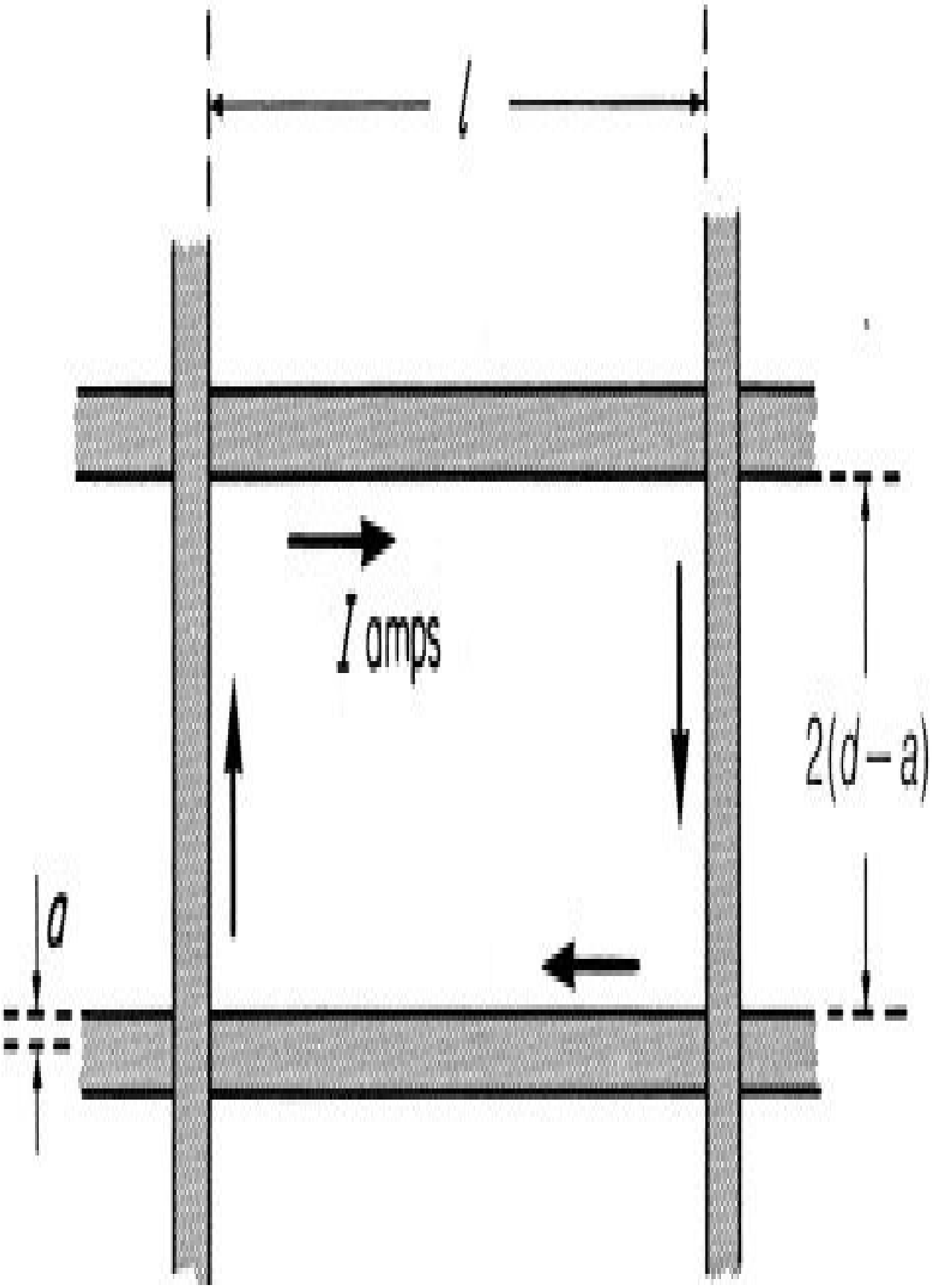
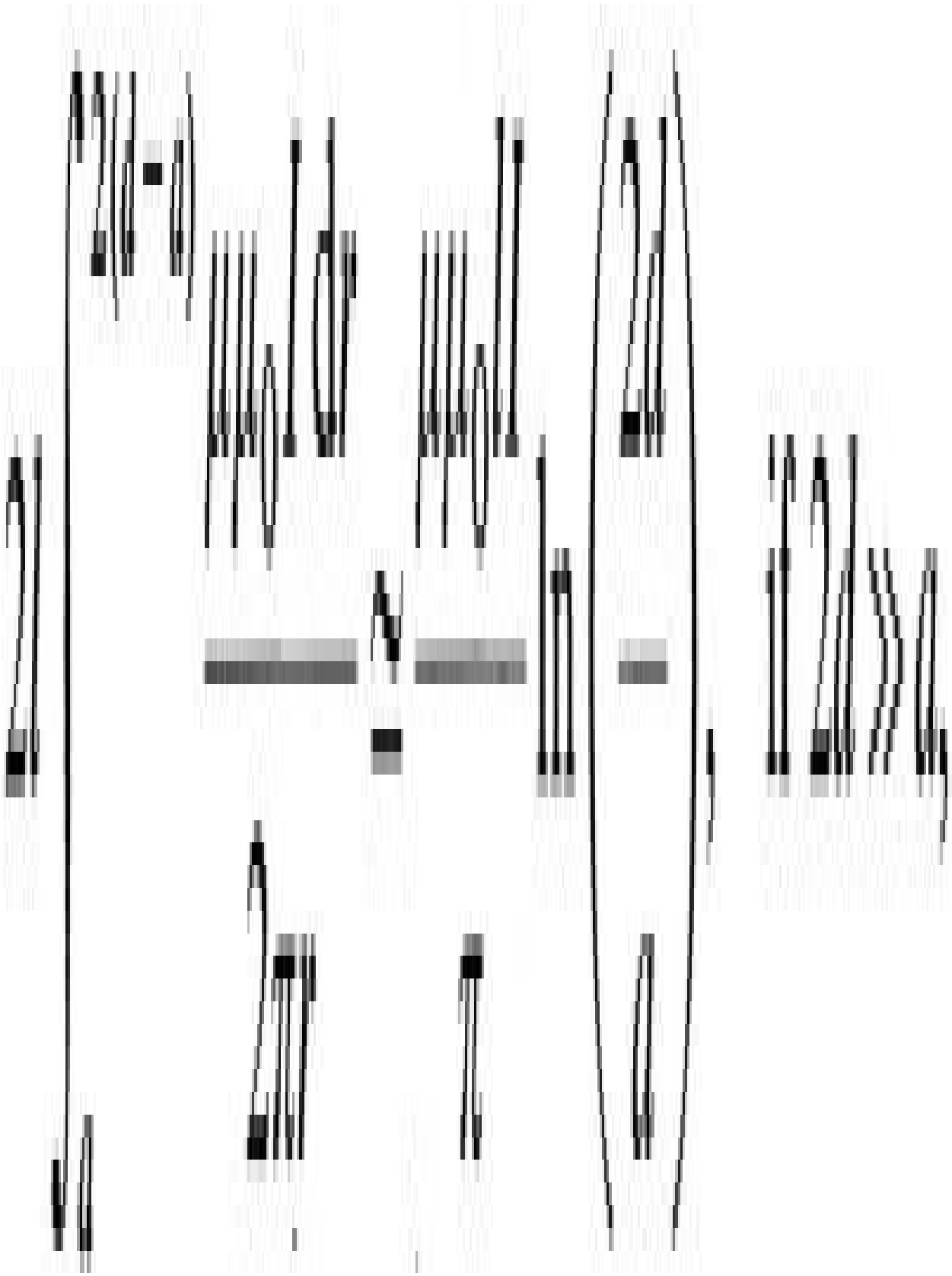

b
2a

Figure 9.5. To find the self-inductance per unit length of a transmission line, a closed circuit is made by connecting the conductors at two points along the line.


l
I amps
2(d − a)
a

magnetic flux through the circuit is

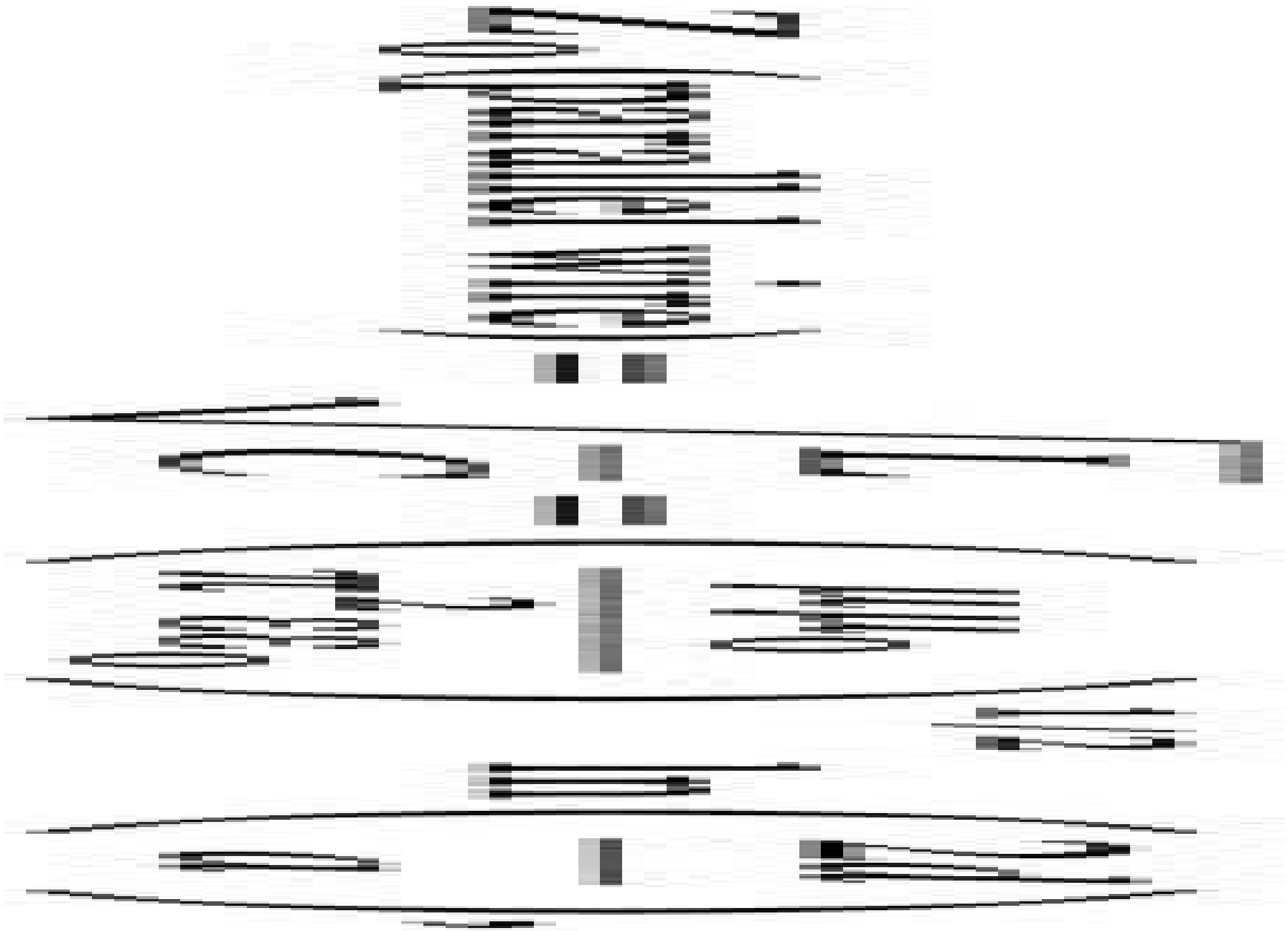


and the self-inductance per unit length is

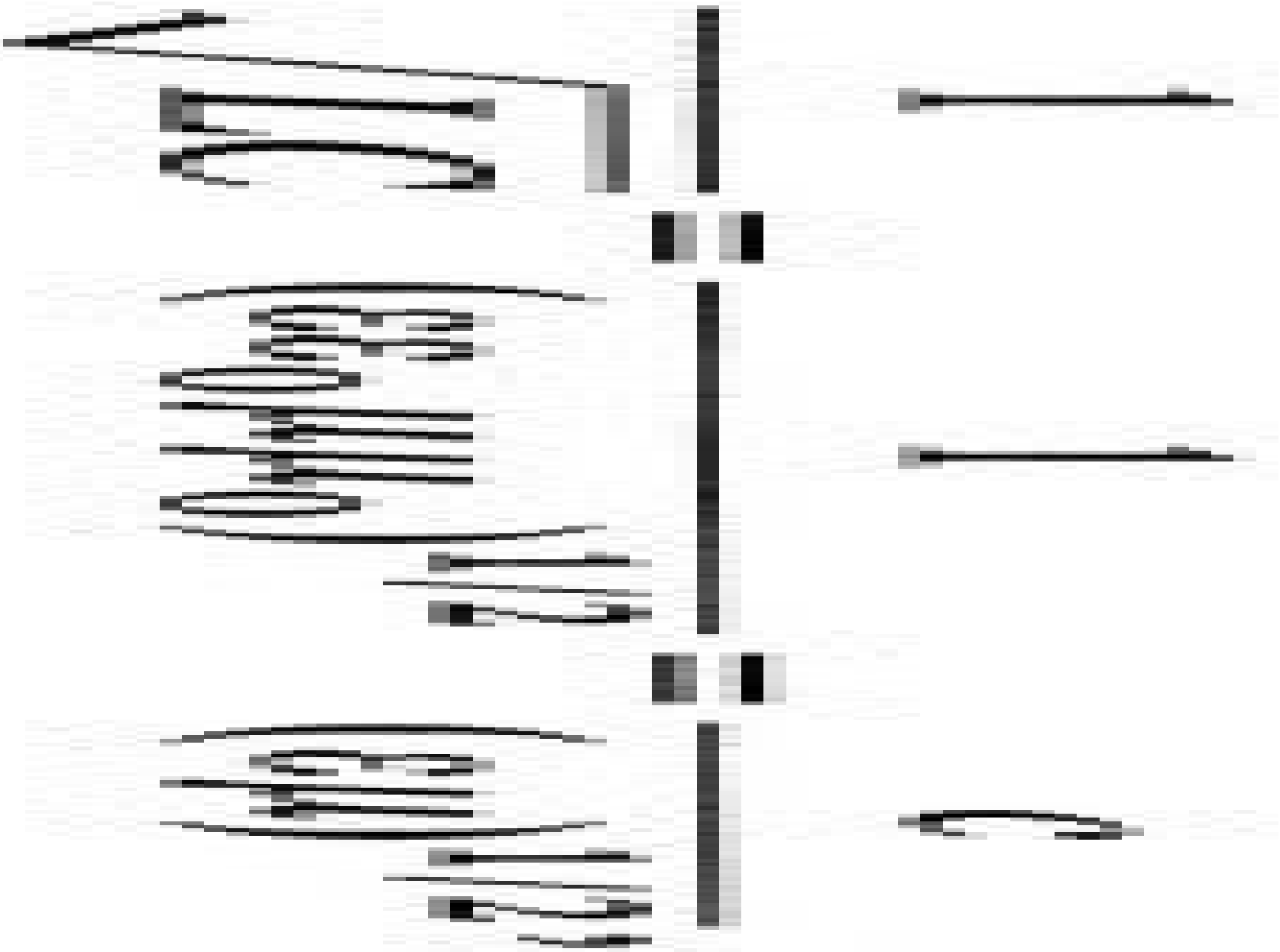
(9.8)

(There is actually a contribution to the self-inductance from the flux passing through the conductors. This is negligible when 2d » a.) From Equations (9.7) and (9.8) we find that the characteristic impedance of the line is

(9.9)



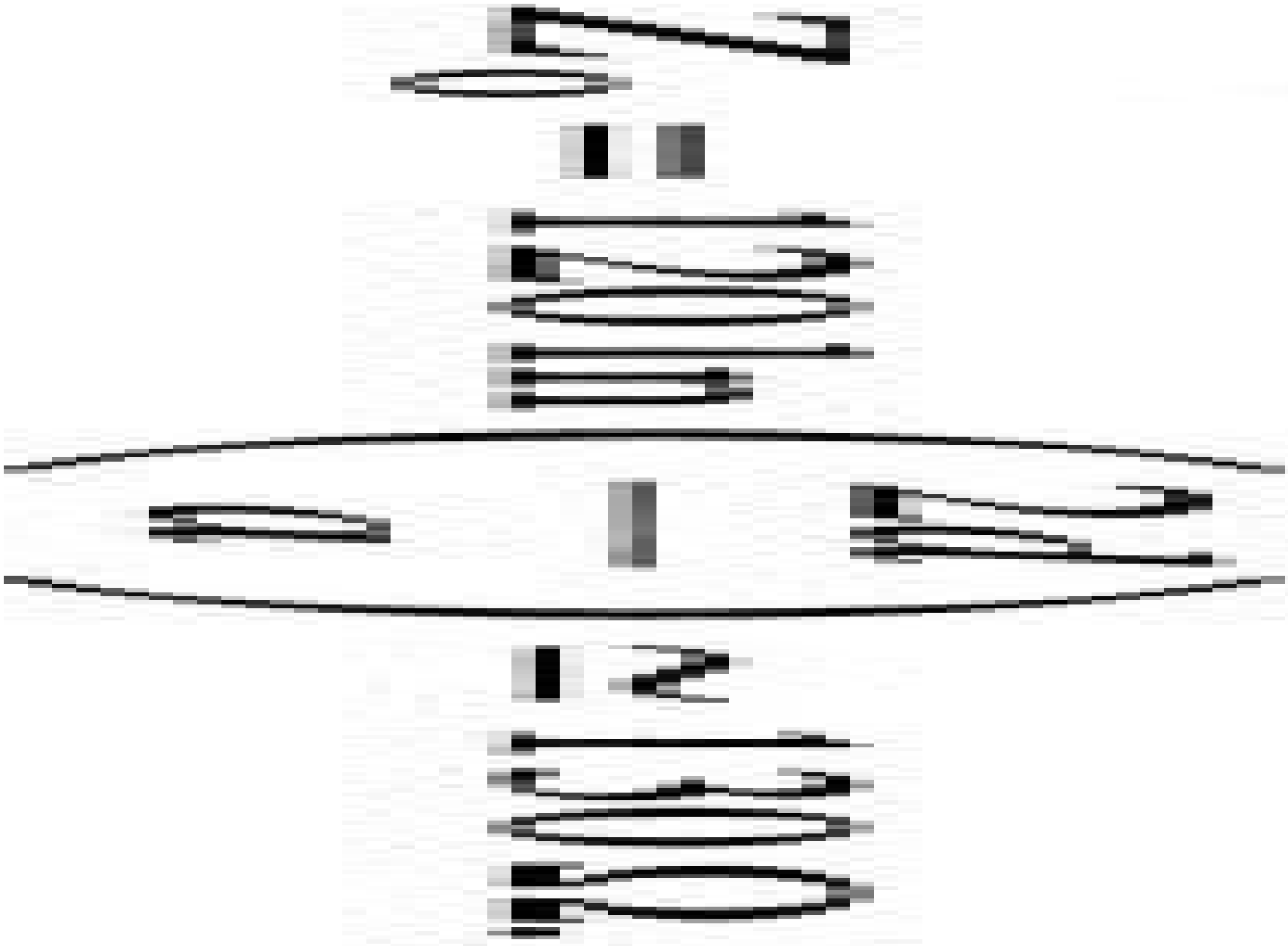
and the speed of propagation of signals



a constant, independent of the dimensions of the line. As we shall find in Chapter 11 when we come to discuss radiation, the constant c is the speed of light in vacuo, and $c/(\varepsilon\mu)^{1/2}$ the speed of light in the medium in which the line is embedded.

It turns out that the velocity of propagation is equal to the speed of light for any lossless line, whatever the size and shape of its conductors.

Parallel wires, twisted to minimize their external electric and magnetic fields, are used to carry signals over fairly short distances, for example to connect a number of computers in the same building. The wires are each insulated, but the whole transmission line is not completely embedded in dielectric material. The wires are also close together, so the approximation that $2d \gg a$ used in deriving Equations (9.7) and (9.8) is not valid. However, we can get a rough idea of the characteristic impedance by using this approximation and by assuming that the lines are in air, for which we may take $\varepsilon = \mu = 1$. Putting $2d/a = 3$, Equation (9.9) then gives

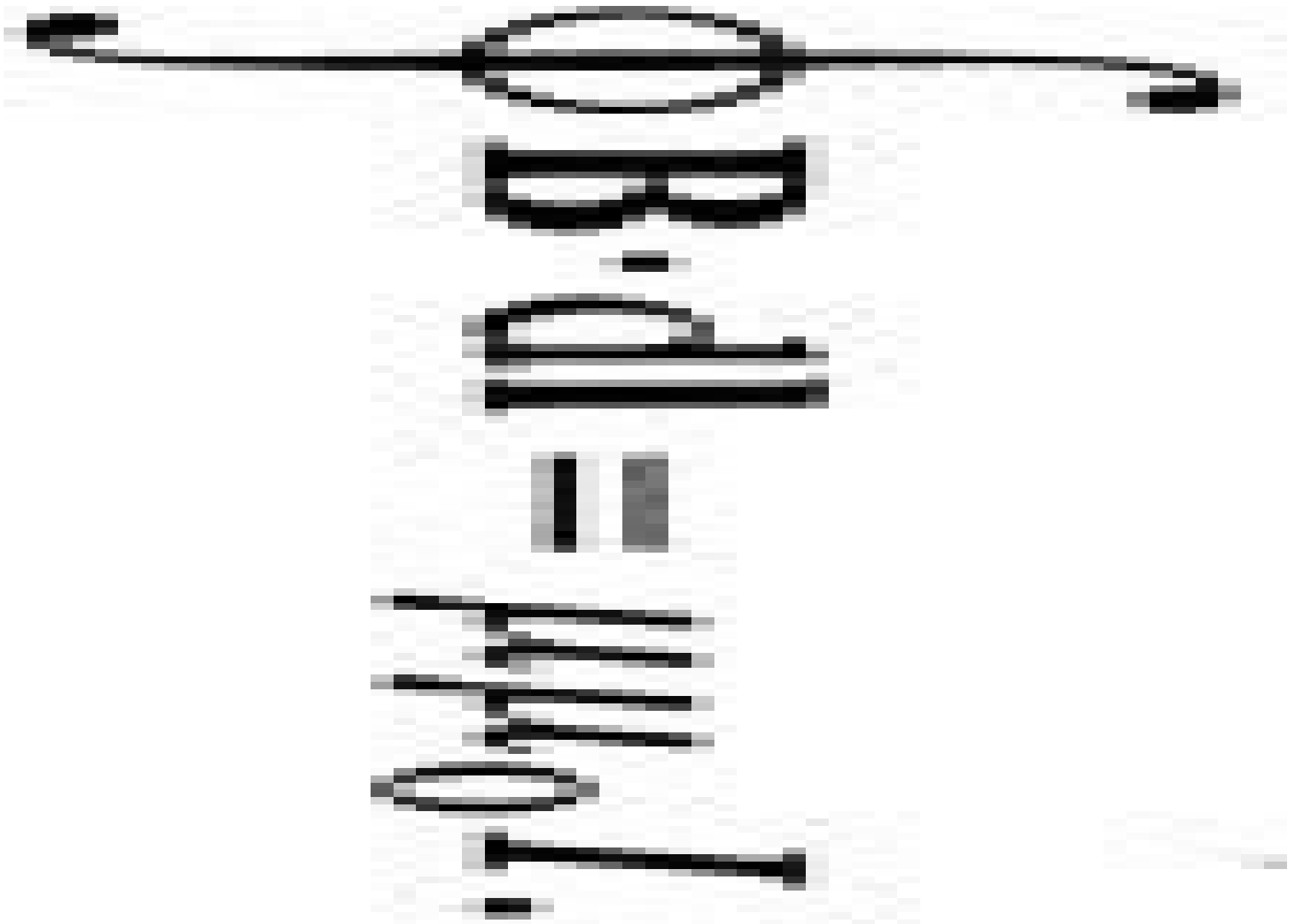


9.2.2 The coaxial cable

The coaxial cable consists of two coaxial cylindrical conductors with cross-section as in Figure 9.6. The conductors are separated by a material of relative permittivity ε and relative permeability μ. The capacitance per unit length has already been found to be (see section 2.3.2)

(9.10)

Suppose that the conductors each carry a current of I amps. Applying Ampère's law to the circuit of radius r in Figure 9.6.



Because the cable has cylindrical symmetry, the magnitude B of the field depends only on r, and

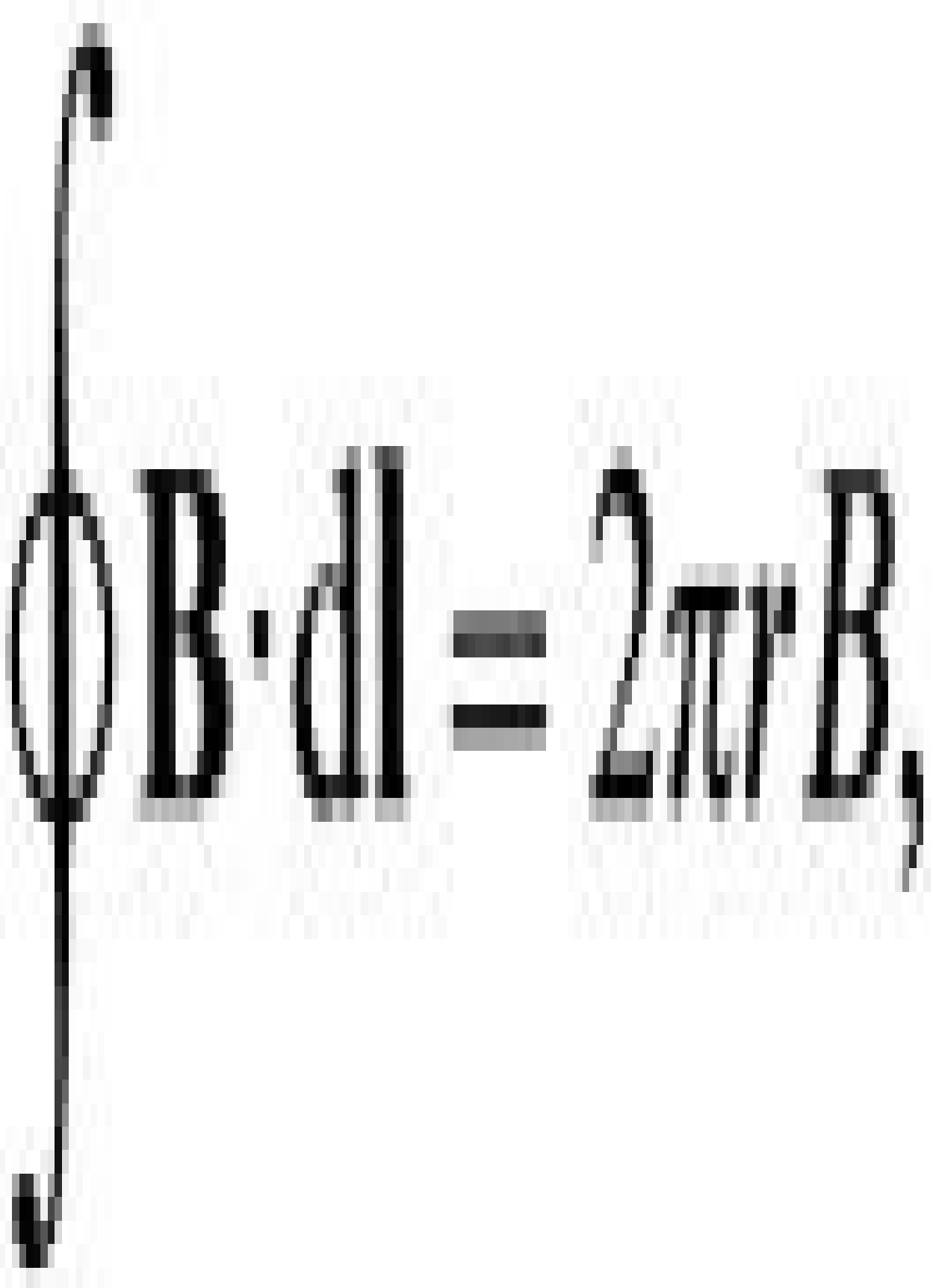
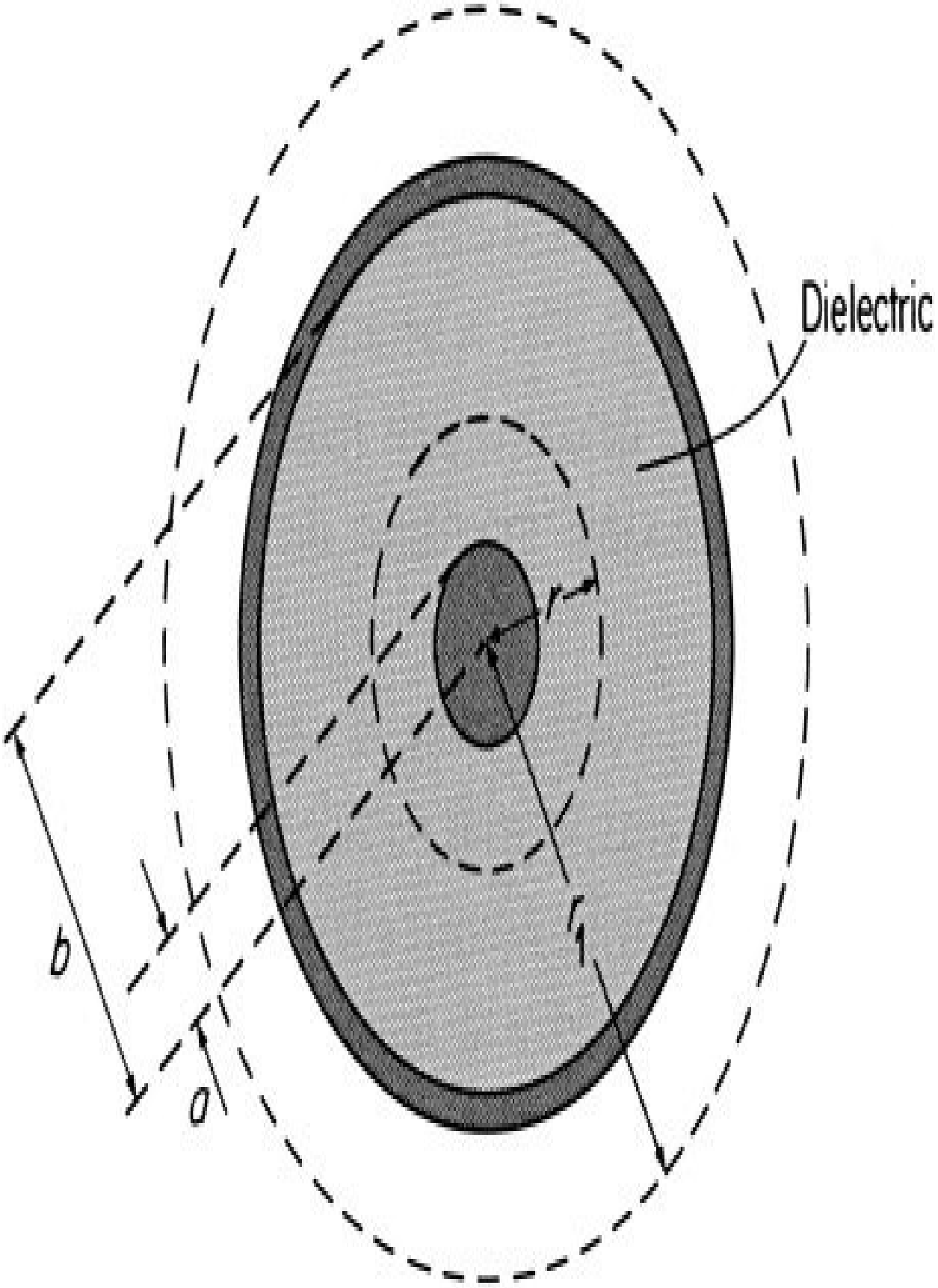
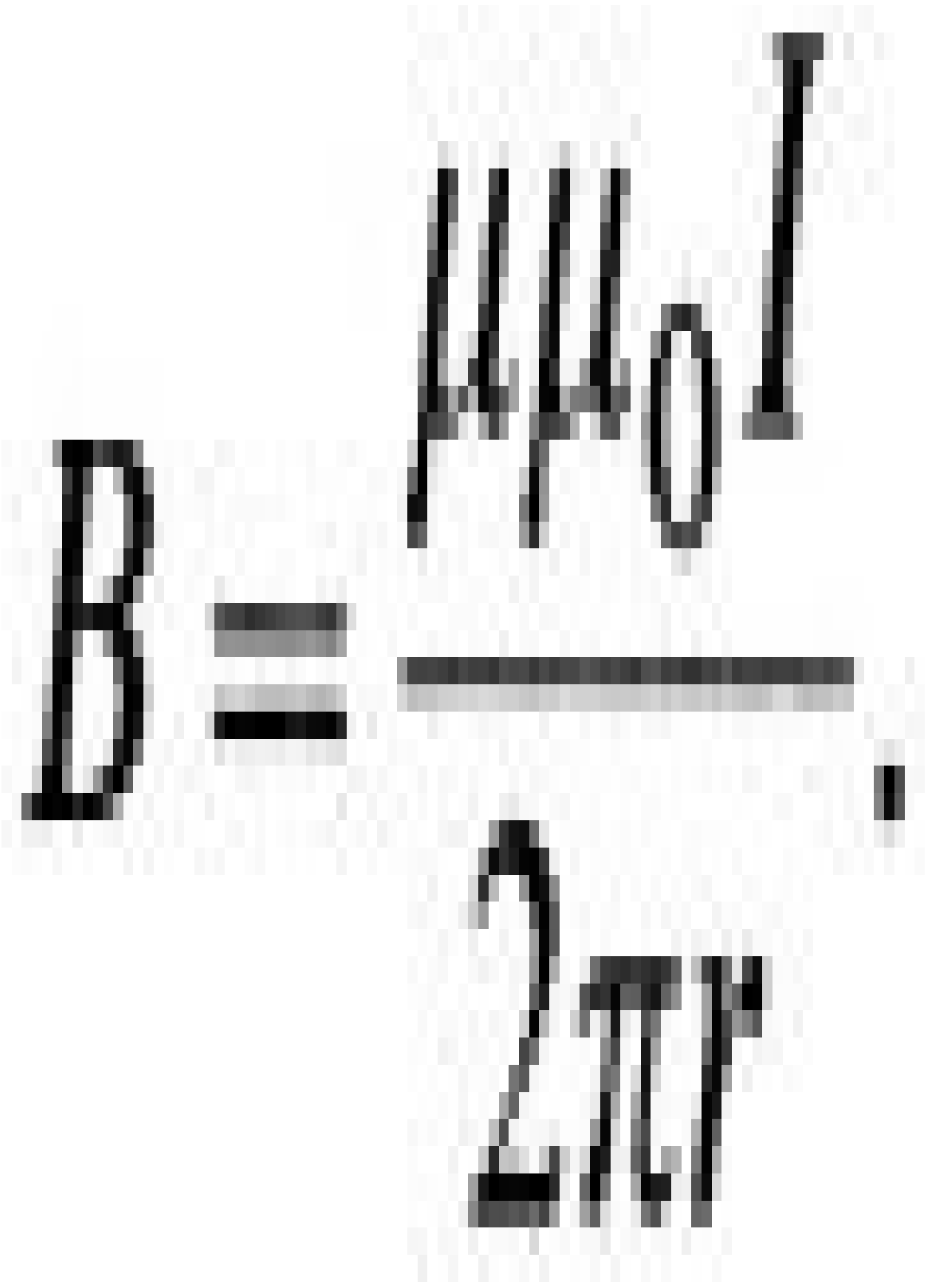

$\oint B \cdot dl = 2\pi r B_\phi$

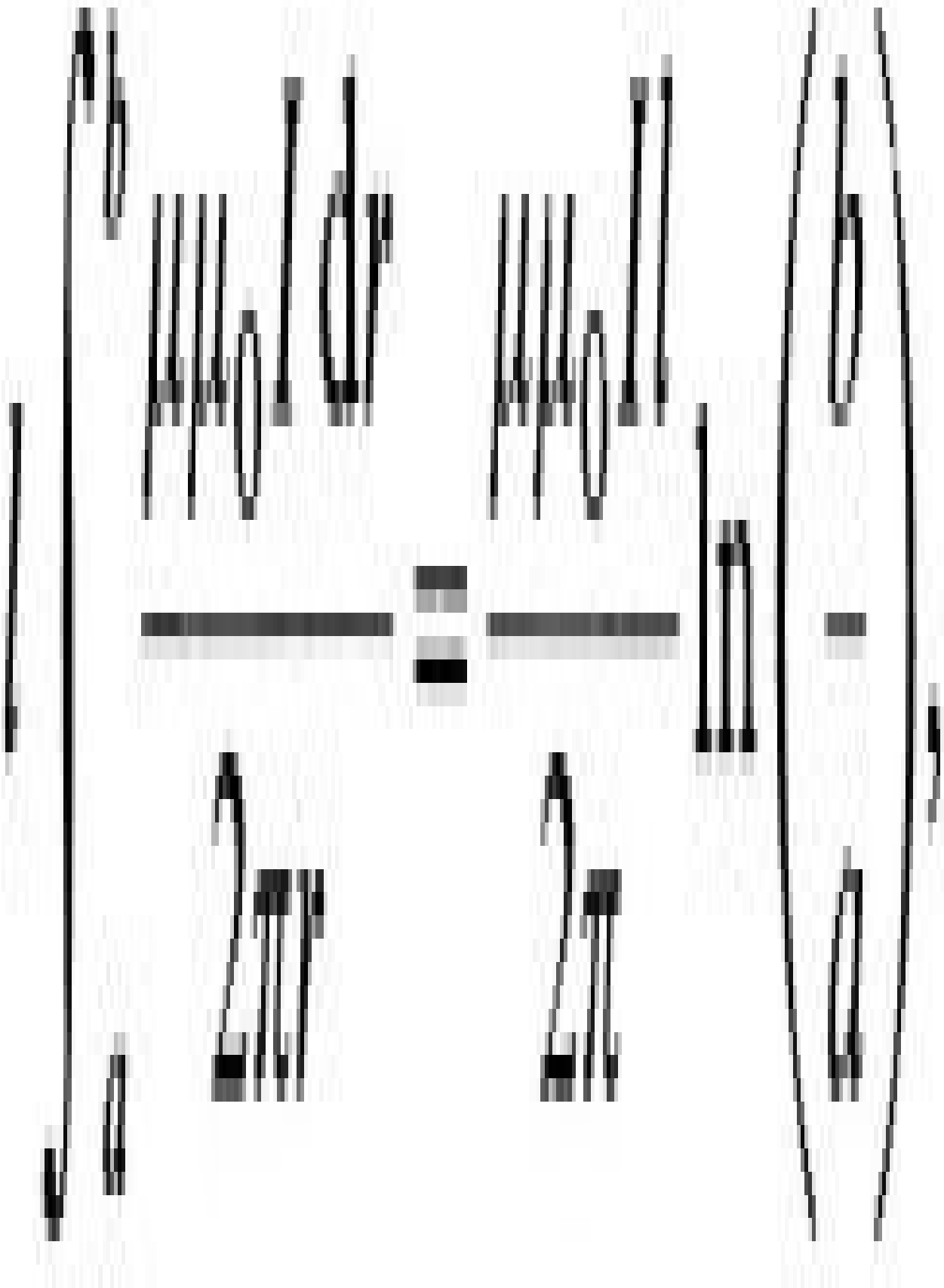
Figure 9.6. Cross-section of a coaxial cable.


Dielectric
b
a

leading to



The flux of the field B through a closed circuit formed by joining the conductors at the end of a section of cable of length l(l » b) is

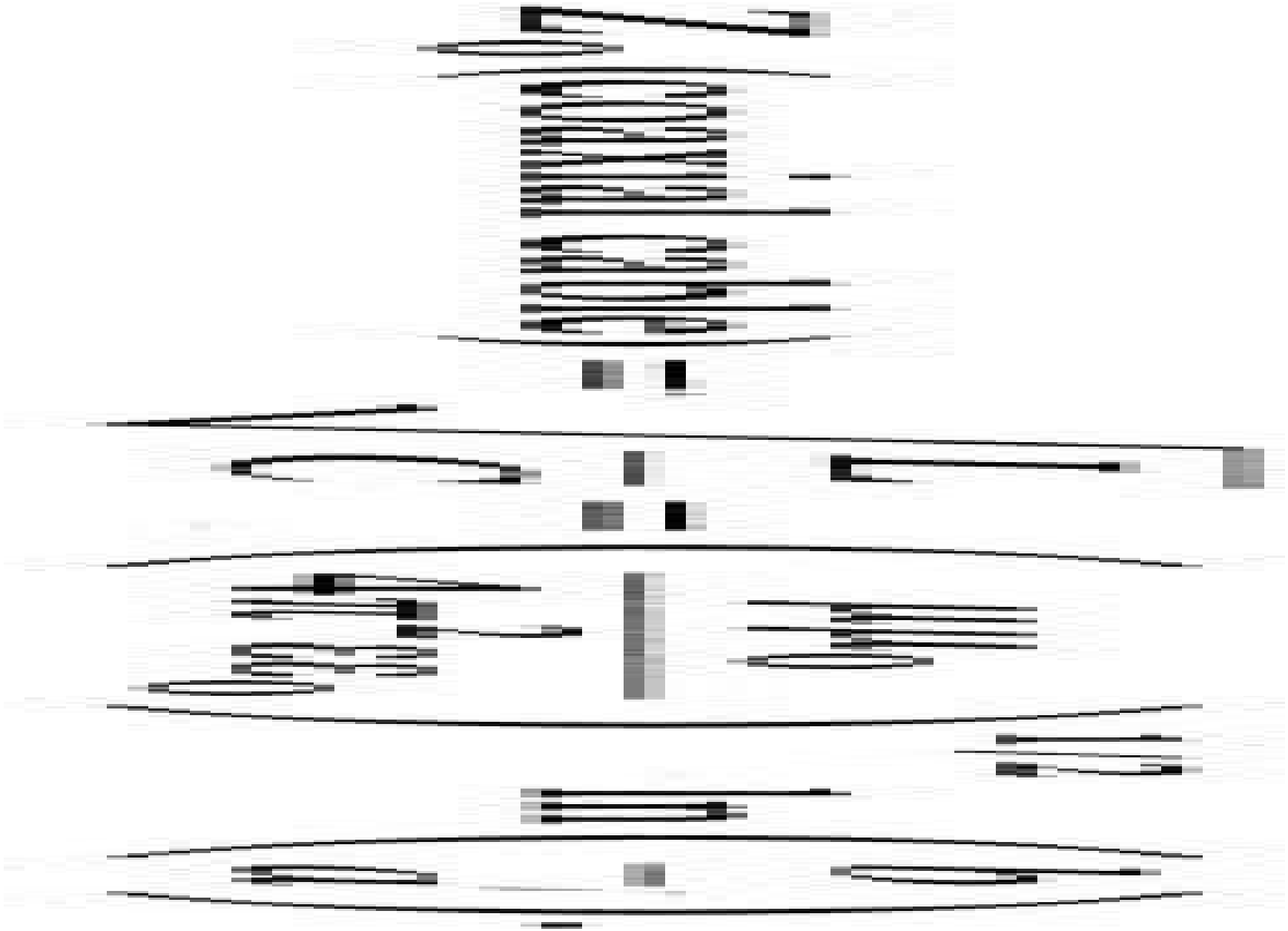


and the self inductance per unit length is

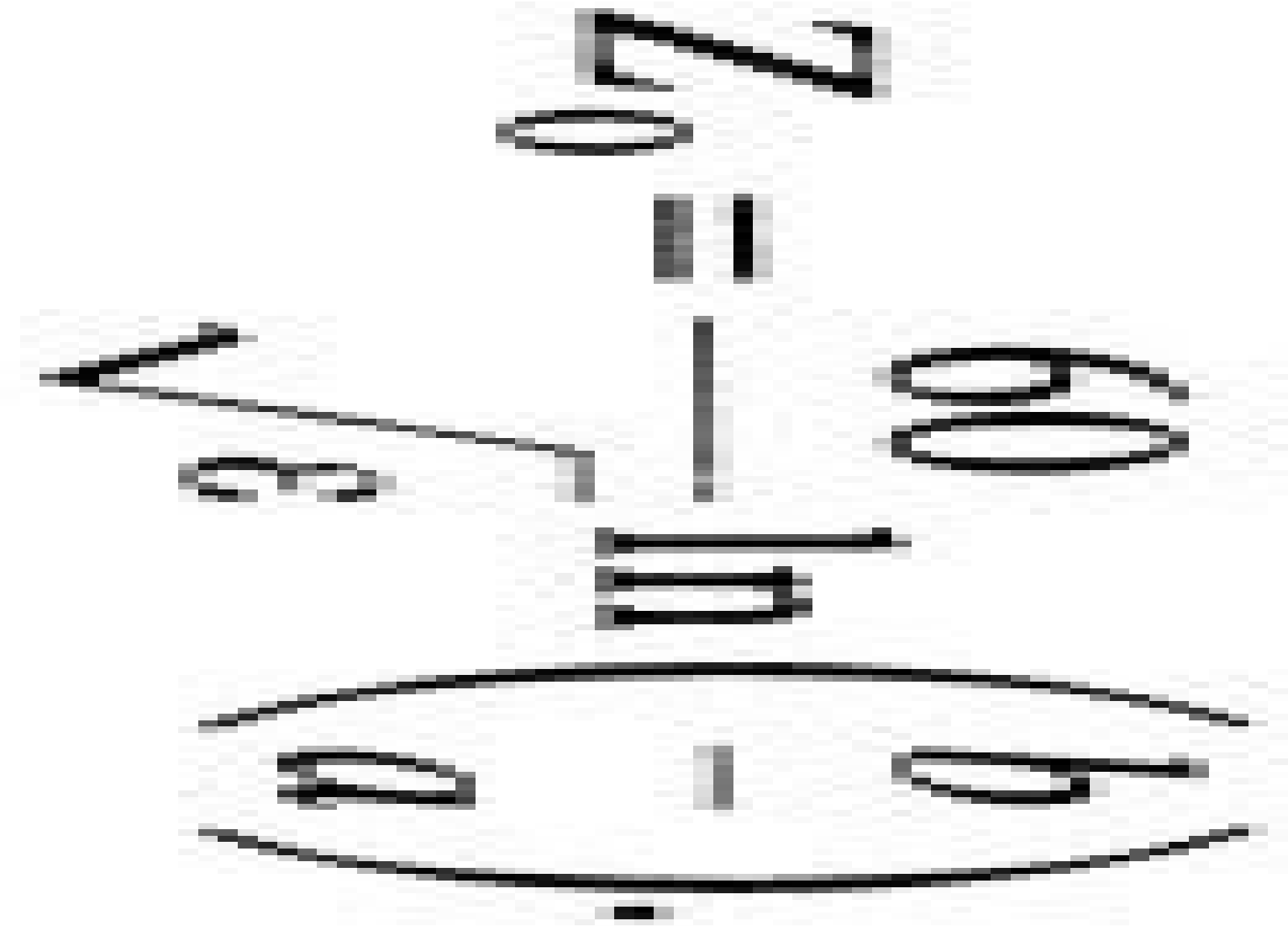
(9.11)

From Equations (9.10) and (9.11) we check that the speed of propagation is again $c/(\varepsilon\mu)^{1/2}$, and obtain for the characteristic impedance of the cable

(9.12)



Usually the insulator between the conductors of a coaxial cable is a solid with relative permeability very nearly unity, and we may write



For example, if the insulator is polyethylene with relative permittivity 2.3 then $Z0 = 50\Omega$ for $b/a = 3.5$. Fifty ohms is often chosen as the characteristic impedance of coaxial cables for laboratory use. Many items of commercially available electronic equipment are standardized to have input and output impedance of fifty ohms. This allows signals to be transmitted along 50Ω cable from one piece of equipment to another without reflections, as is explained in section 9.3. The cable linking a television set to its antenna, however, has a characteristic impedance of 75Ω. This is because the antenna has an effective impedance close to 75Ω, and detects signals most efficiently when connected to a transmission line of the same impedance: the properties of antennas are discussed in section 13.3.

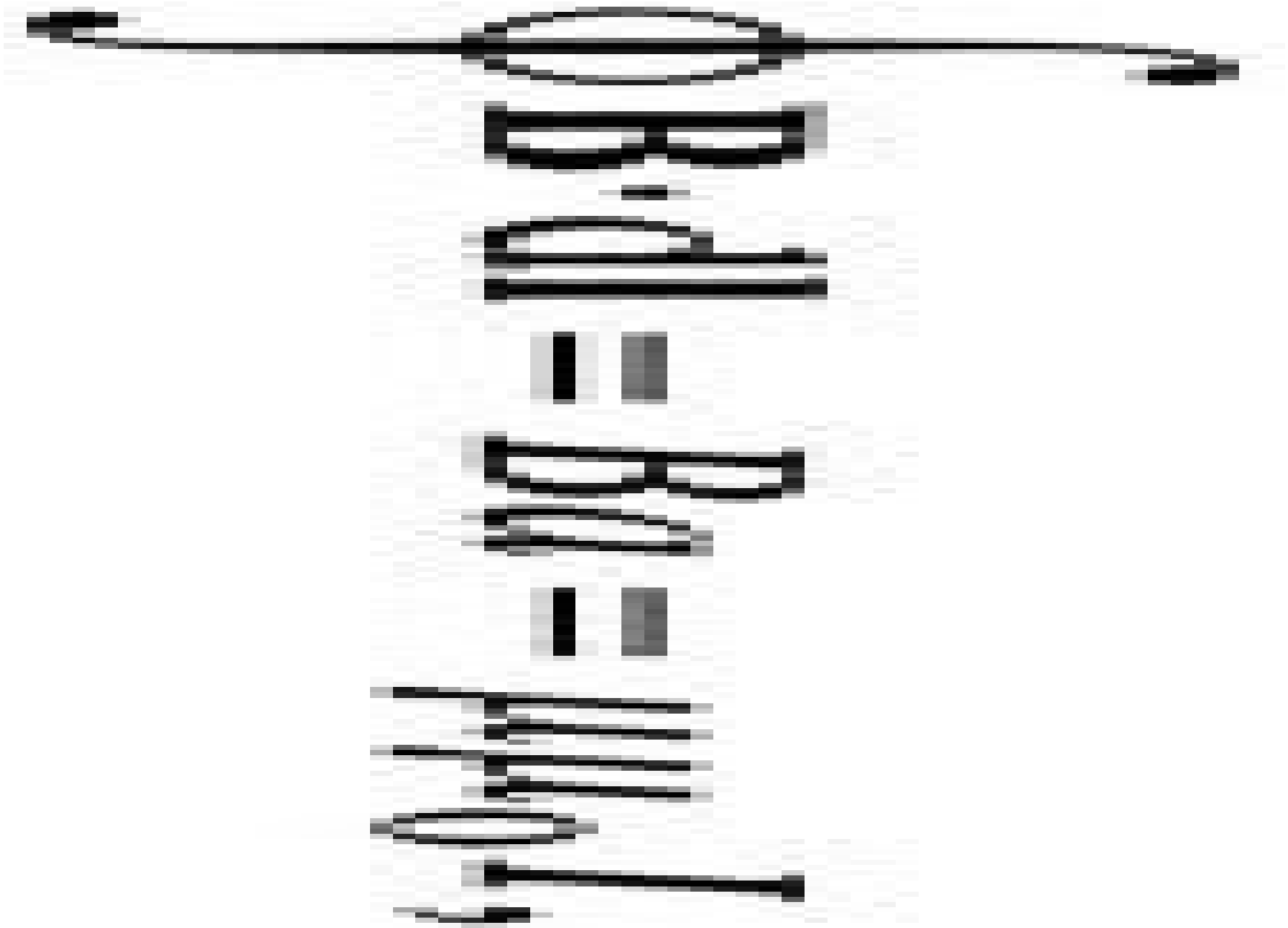
The outer conductor of a coaxial cable is normally maintained at earth potential. There are then no electric fields outside the cable. Nor are there any magnetic fields outside it. Applying Ampère's law to the circuit of radius r1 in Figure 9.6, $B.dl = 0$, since equal and opposite currents flow in the two conductors of the cable: hence by symmetry $B = 0$ outside. The confinement of the fields within the outer conductor gives the cable a two-fold advantage over the parallel wires. In the first place, the absence of external magnetic fields means that there can be no unwanted inductive coupling between the cable and other circuits. Secondly, there is no radiation from a coaxial cable. In Chapter 11 we shall learn that energy must be radiated from a pair of parallel wires carrying high frequency currents. This leads to attenuation of signals even when there are no resistive losses.

9.2.3 Parallel strip lines

Figure 9.7 shows the cross-section of a transmission line made of two parallel strips of conductor attached to the opposite faces of a slab of dielectric material. Neglecting edge effects*, as we did in deriving the capacitance of a pair of parallel plates in section 1.5.5, the capacitance per unit length is

(9.13)

For a strip carrying current I, the magnetic field in the gap between the strips is found by applying Ampère's law to the dotted contour in Figure 9.7. Except near the edge of the strip, there is no field outside the transmission line, since a contour right round the whole line encloses no current. Along the dotted contour, the line integral



and the magnetic field B is uniform within the transmission line, having a magnitude

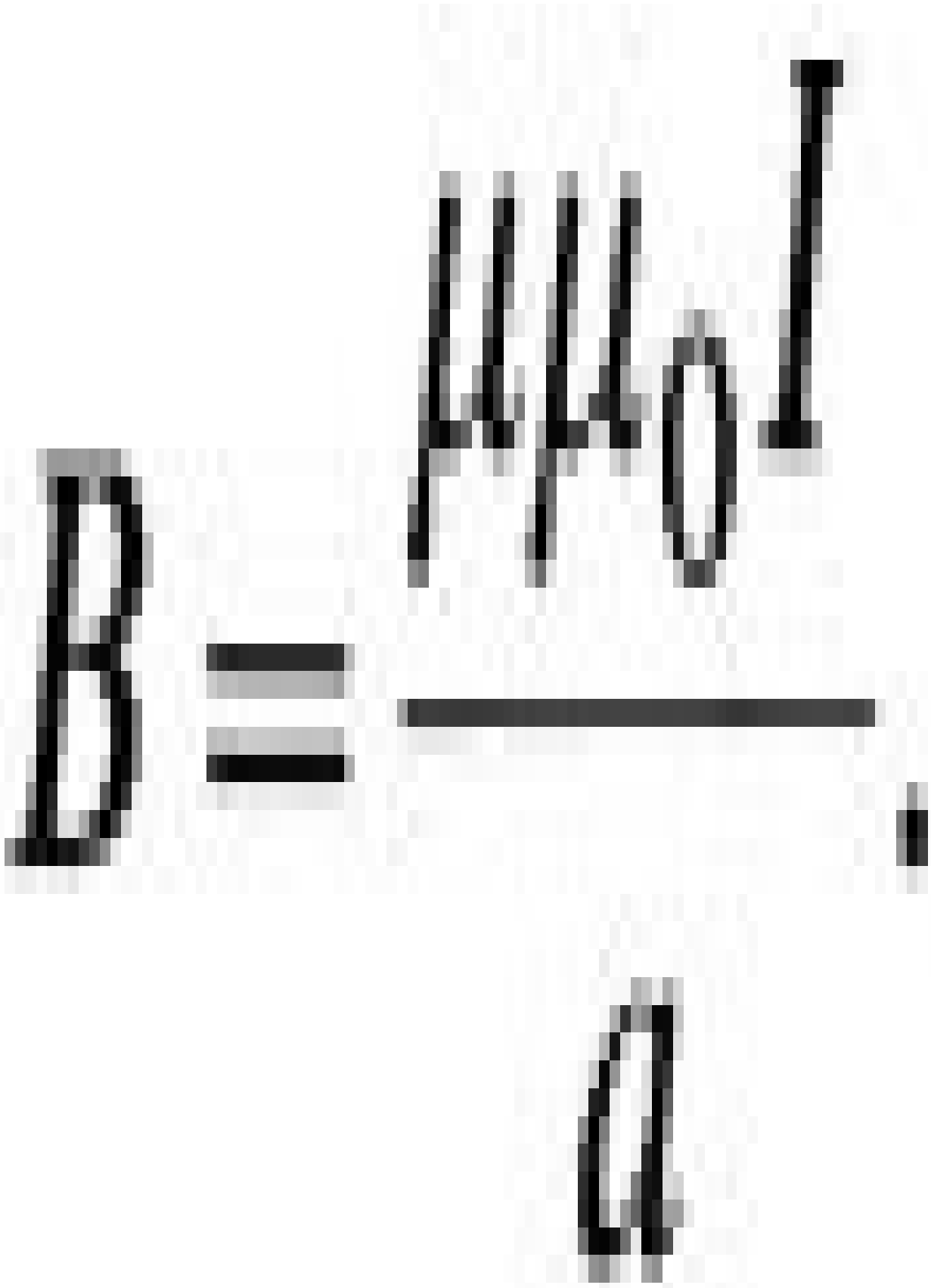
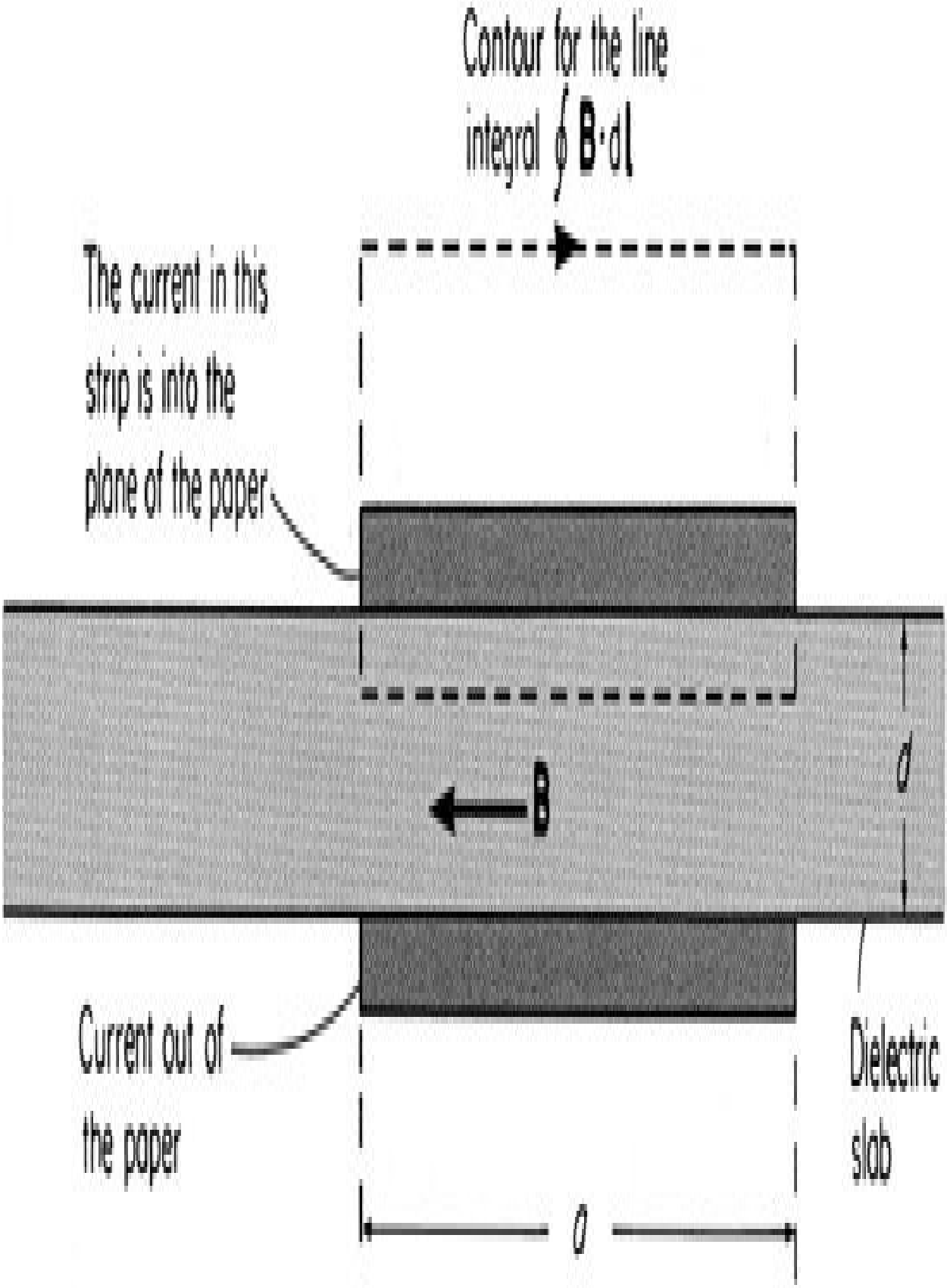


Figure 9.7. Cross-section of a strip line.


Contour for the line integral $\oint \mathbf{B} \cdot d\mathbf{l}$
The current in this strip is into the plane of the paper
B
d
Current out of the paper
Dielectric slab
a

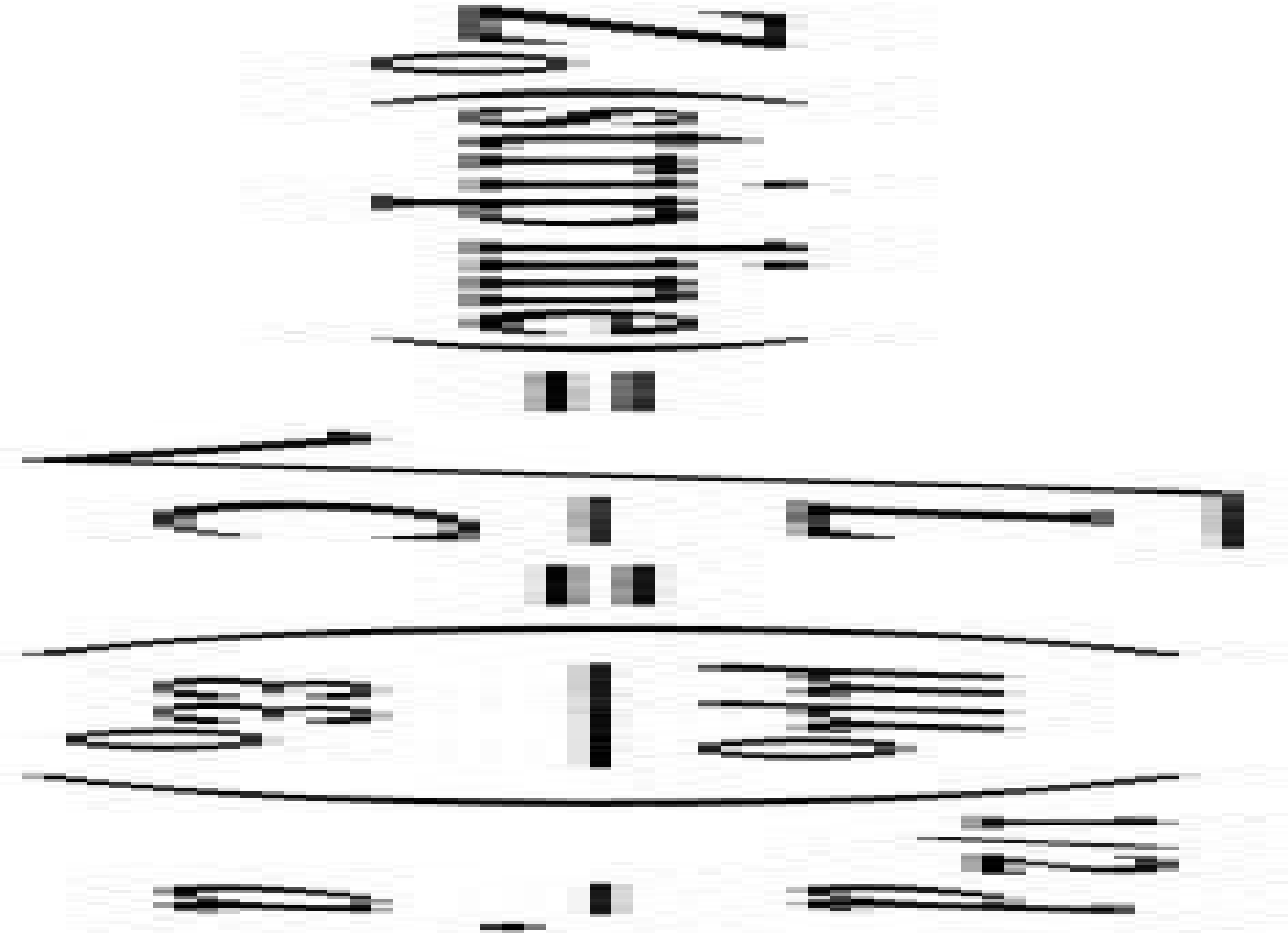
When a closed circuit is formed by connecting the ends of a length l of the line, a flux, μμ0Idl/a threads it, and the self-inductance per unit length is therefore

(9.14)

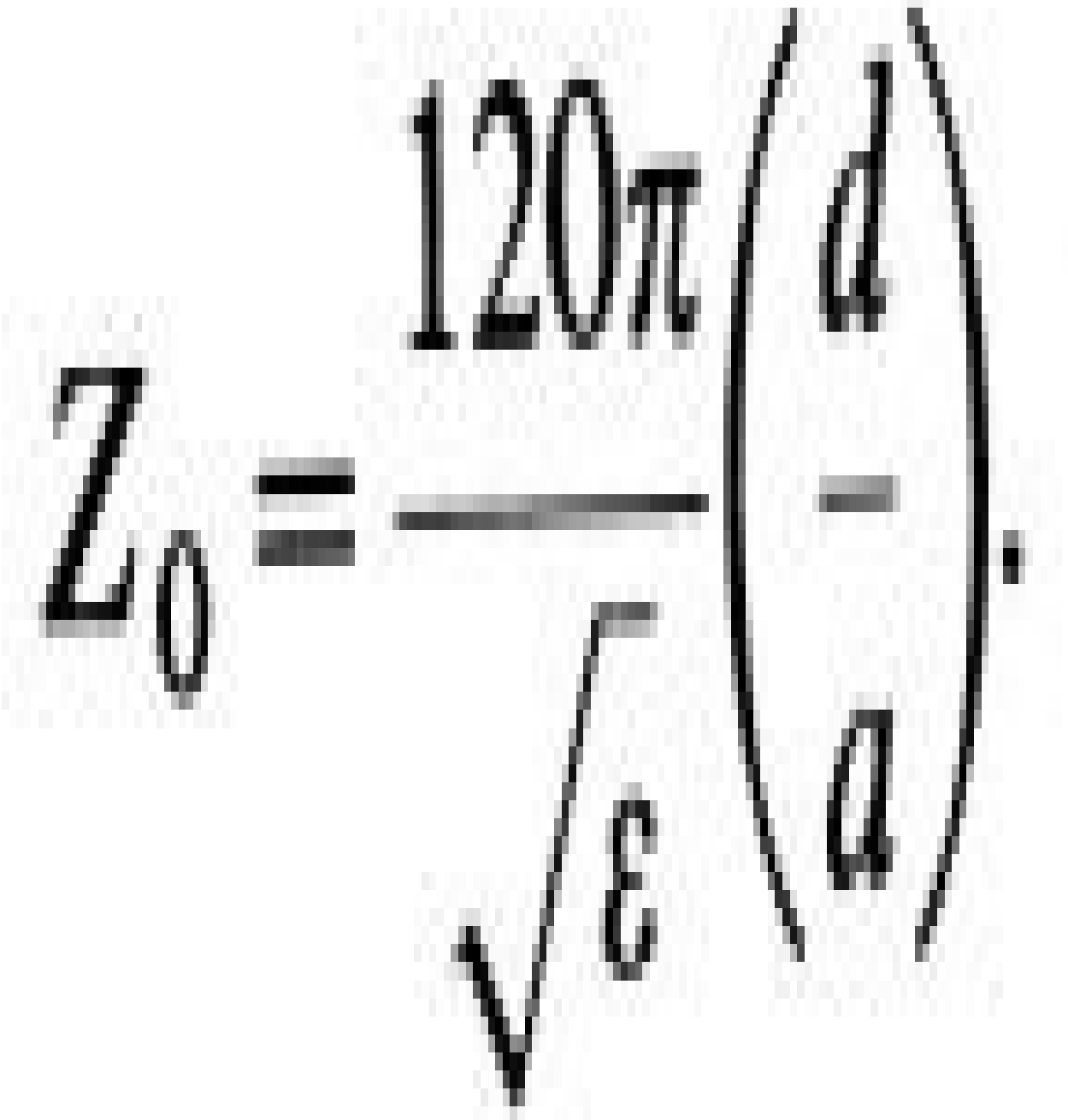
Once again the speed of propagation 1 / is $c/(\varepsilon\mu)^{1/2}$.

The characteristic impedance of the line is

(9.15)



If the strips are separated by a material with relative permittivity ε and relative permeability unity, this expression reduces to



A property of the strip line which is often useful is that the characteristic impedance varies directly as the ratio d/a of spacing to width, in contrast to the slow logarithmic dependence on dimensions for both the parallel wire and coaxial lines. Low impedances can be achieved by choosing d « a. It is particularly easy to connect together strip lines of different characteristic impedances, simply by fixing different widths of conductor to the same slab of dielectric material.

Strip lines are used to transport signals around the printed circuit boards on which electronic components are mounted. One side of the board usually carries the earthed strips, and the other all the signal voltages. Although strip lines are not totally enclosed like coaxial cables, the fields outside the lines are much less than those outside parallel wires, and ‘cross-talk’ between neighbouring strips on a board can be kept to acceptably low values at frequencies up to several hundred megahertz. Modern communications circuits operate at even higher frequencies in the GHz (10^9 Hz) range, and these too make use of strip lines. At these frequencies the radiative coupling between different components is important, and great care must be exercised in choosing the layout of the whole circuit.

Metallic strips are used in a somewhat different way to carry signals within integrated circuits. In this case it is not usually possible to place the two strips opposite one another because of the three-dimensional structure of the circuit. Instead the strips are laid side by side, sometimes enclosed in a sandwich of dielectric materials of different relative permittivities. The field patterns around such lines are complicated, and must be calculated using numerical techniques in order to estimate the characteristic impedances and interaction with other lines. Integrated circuits are also being used at ever-increasing frequencies, and it is becoming more and more important to have accurate calculations of the properties of their strip lines.

9.3 REFLECTIONS

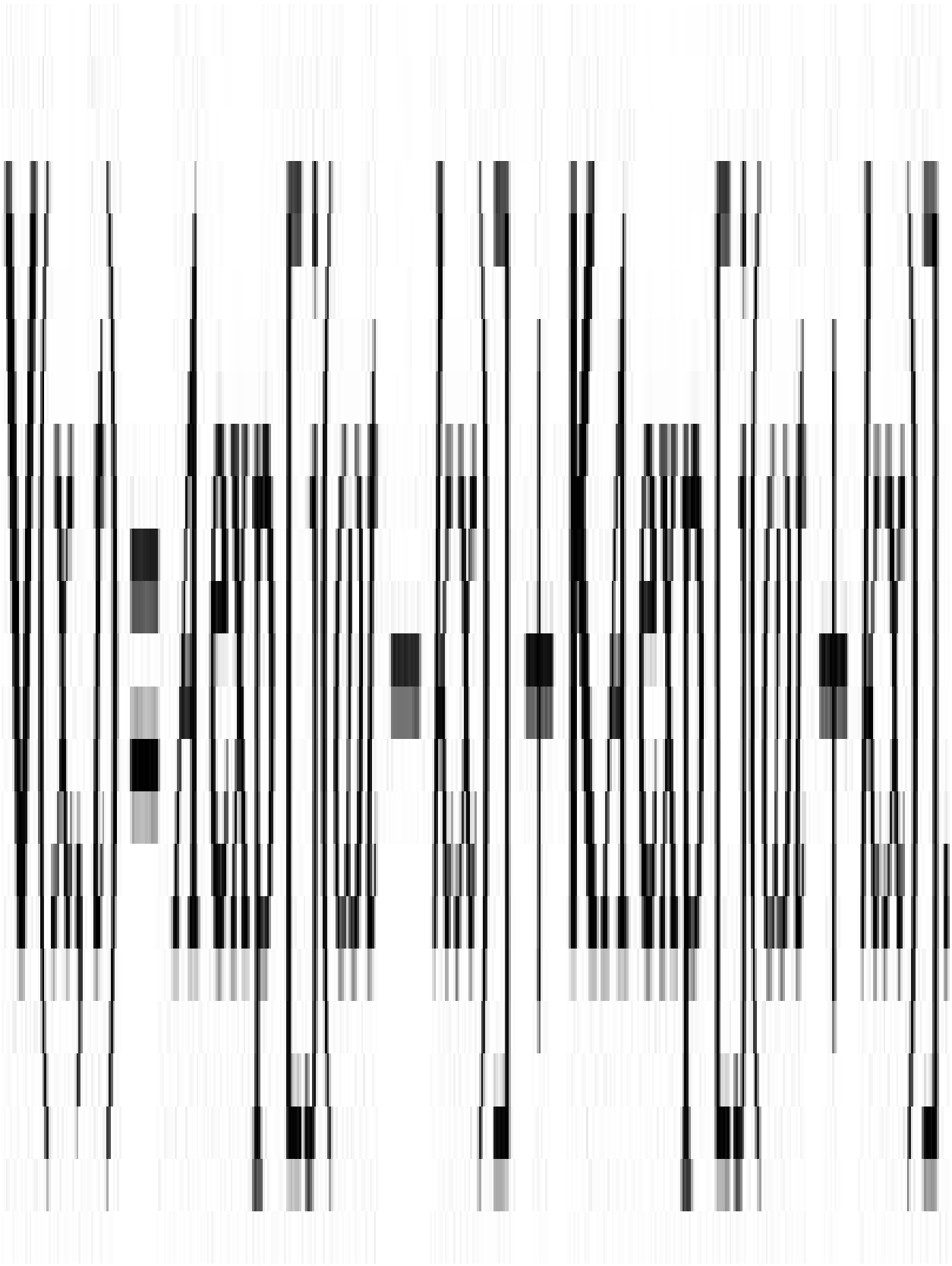
If a transmission line is terminated by a resistance equal to its characteristic impedance Z0 then an alternating voltage $V_0 \cos \omega t$ applied to the left hand end of the line generates a wave described by Equations (9.5). Rewriting these equations in the complex notation:

$$V = V_0 e^{j(\omega t - kx)}$$

$$I = \frac{V_0}{Z_0} e^{j(\omega t - kx)},$$

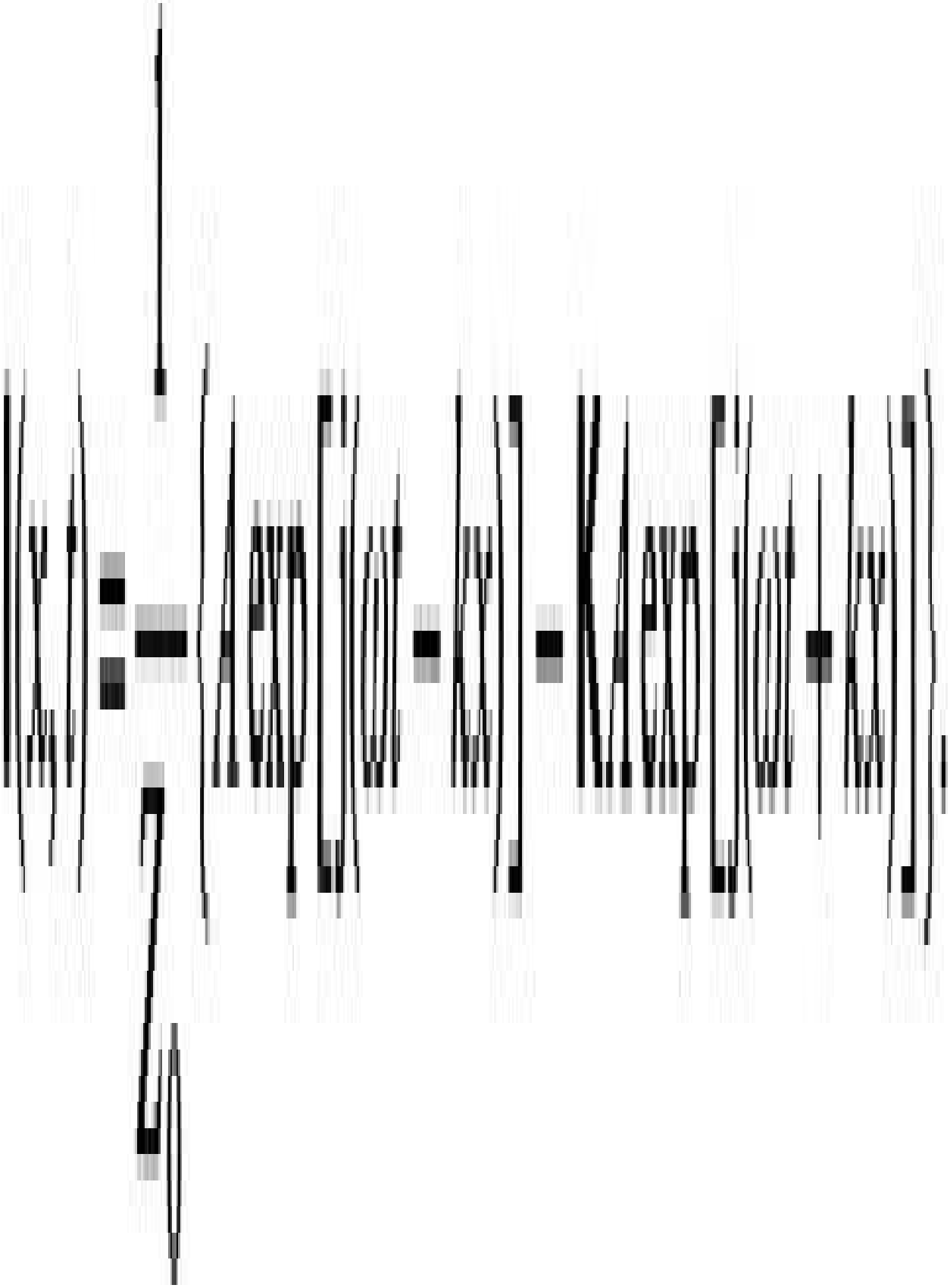
The wave travels forwards from left to right, and is completely absorbed by the terminating resistor. There is no wave travelling in the opposite direction, from right to left. However, if the line is terminated by an impedance Z which is different from the characteristic impedance Z0, the voltage and current no longer have the form given above, since these expressions do not meet the condition V = IZ at the end of the line. In order to satisfy this condition, the solution of the differential equations of the line (Equations (9.1) and (9.2)) must include a term representing a wave travelling back towards the generator. The wave arriving at the terminating impedance is partly reflected, and the line is said to be mismatched. Let us choose the origin of the x-coordinate to be at the end of the line which is terminated by the impedance Z. We write the forward-going wave as A exp [j(ωt – kx)]. This represents a wave travelling from negative values of x towards x = 0, where it makes a contribution A exp (jωt) to the voltage across the terminating impedance. Suppose that the reflected voltage is KA exp (jωt) at x = 0, whence a wave KA exp [j(ωt + kx)] sets out leftwards towards the generator. The quantity K is then the voltage reflection coefficient at the terminating impedance Z. The total voltage on the line is

(9.16)



The expression for the current is similar except that the current reflection coefficient is (—K) (see Equation (9.6)). The current associated with the voltage (9.16) as a solution of the coupled Equations (9.1) and (9.2) is

(9.17)

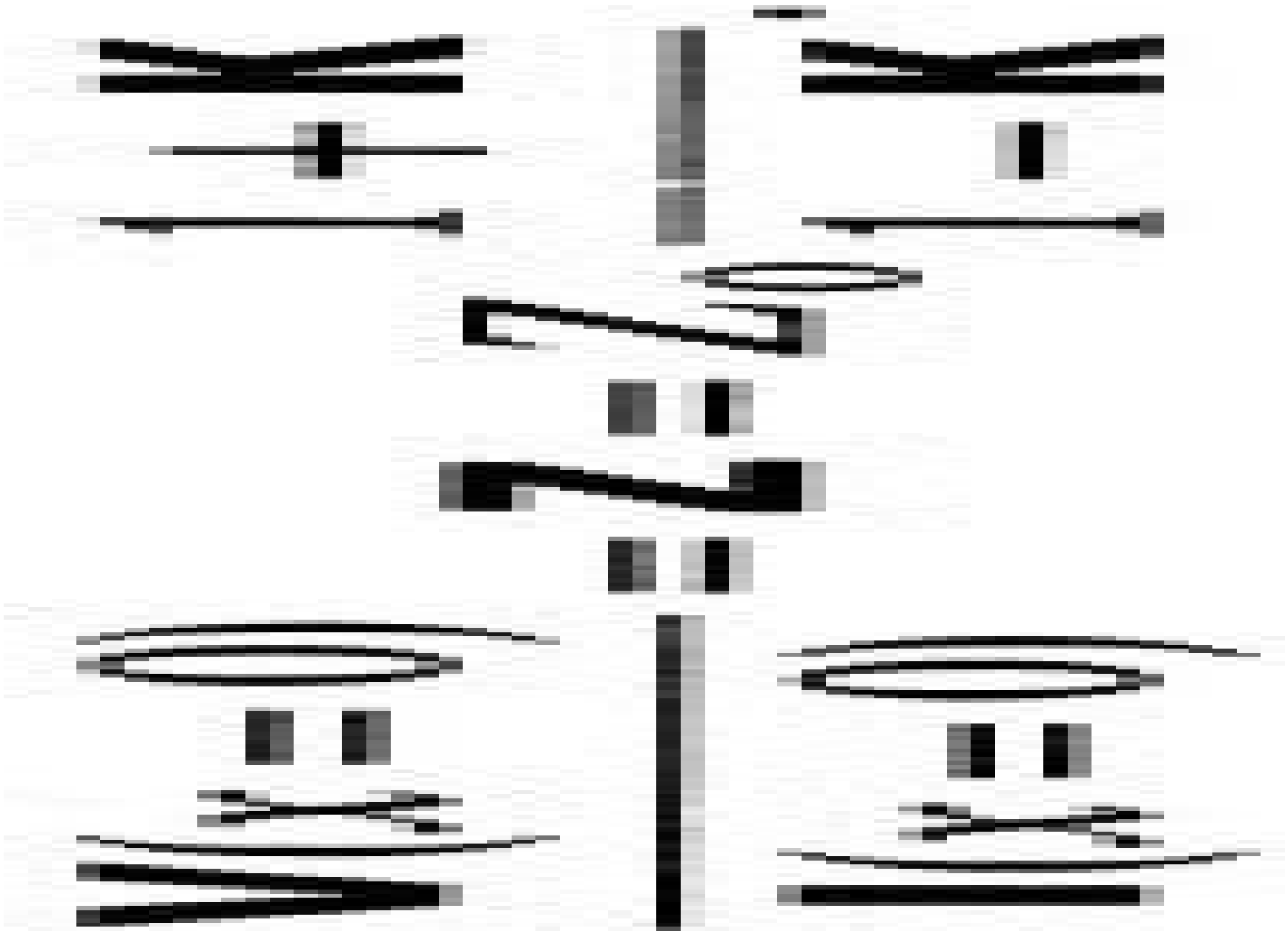


At the end of the line, at $x = 0$, we have

$$V(x=0) = Ae^{j\omega t}(1+K),$$

$$I(x=0) = \frac{A}{Z_0} e^{j\omega t}(1-K).$$

At x = 0 the ratio of voltage to current must equal the terminating impedance Z of the line,



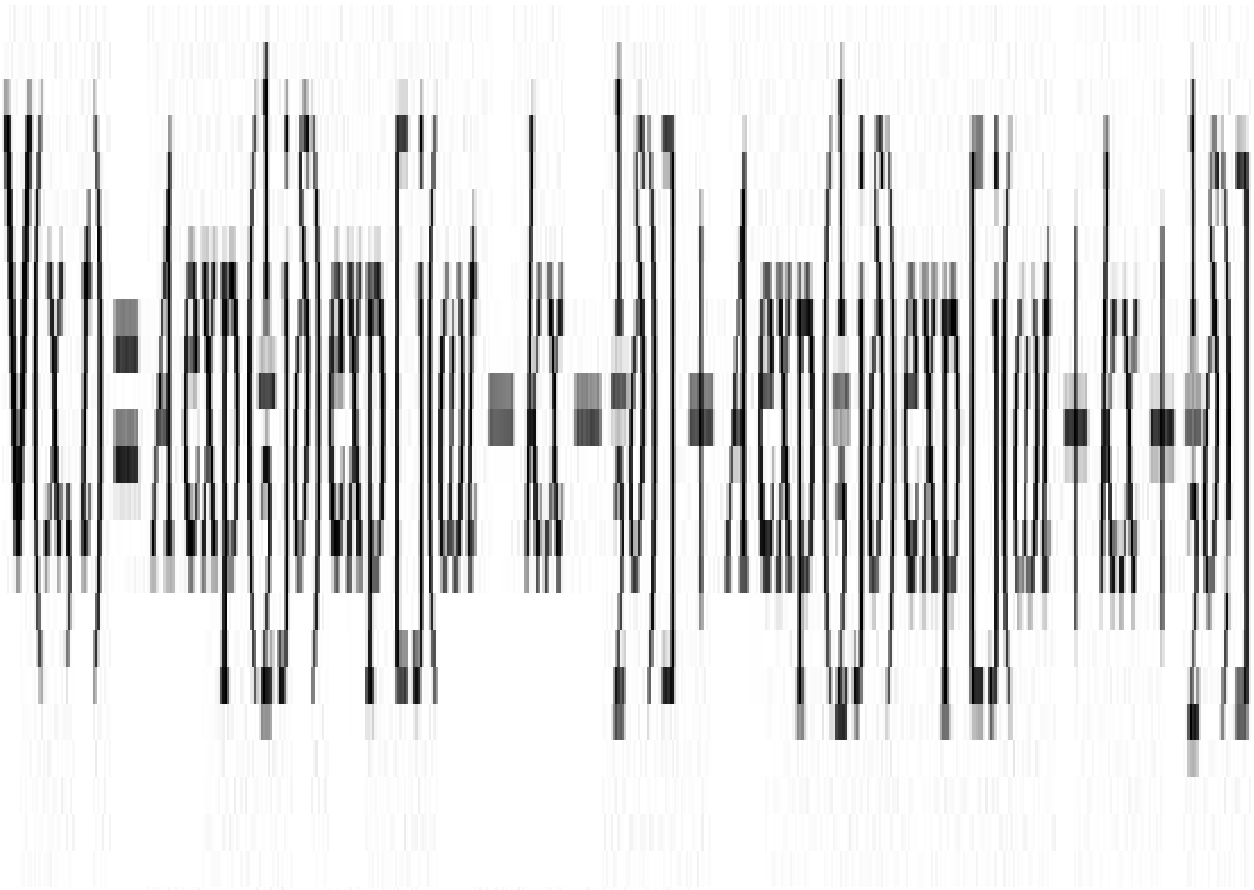
Hence the reflection coefficient

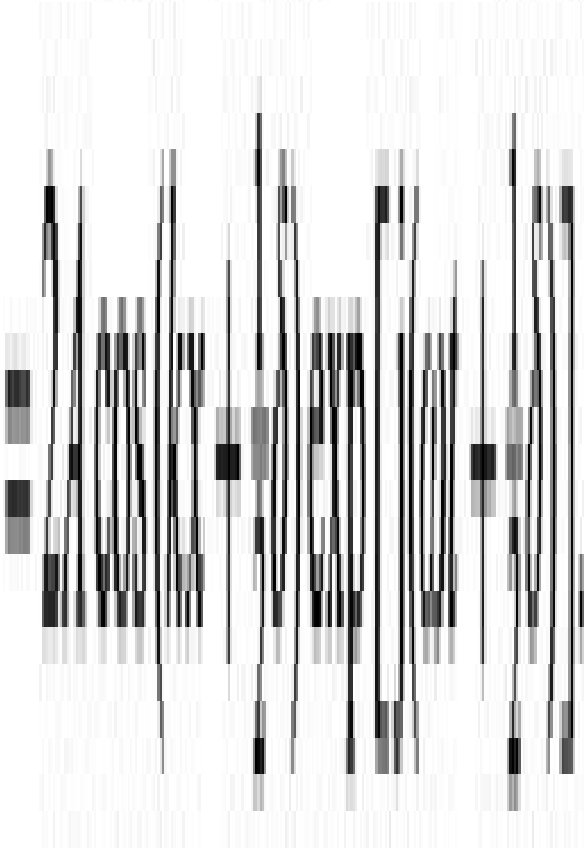
(9.18)

The only way to avoid reflections is to terminate the line with an impedance Z0. All the power carried by the forward-going wave on the line is then absorbed at the termination. At the other extreme, total reflection occurs when the line is open circuited ($Z = \infty$, $K = 1$) or short-circuited ($Z = 0$, $K = -1$). There is no current at the end of an open-circuited line, and no power can be transmitted; on the other hand, no power crosses the end of the shorted line either, since at this point there is no voltage across the conductors. When no power is transmitted, there are standing waves on the line, analogous to the standing sound waves in closed or open-ended pipes. At a node of the standing wave, the voltage across the conductors of the line is zero at all times, and a shorted line thus ends at a node. Nodes are equally spaced along the line with a spacing π/k equal to half the wavelength of the travelling waves. Similarly there are antinodes with a spacing π/k where the voltage has a maximum amplitude, and an open-circuited line ends at an antinode.

No power is transmitted whenever a line is terminated in a purely reactive impedance. The reflected wave has the same amplitude as the forward wave, and there are again standing waves on the line. However, at the end of the line the voltage across the terminating impedance is not in phase with the current. This means that there must be a phase difference between the forward and reflected waves at the end of the line. If this phase difference is δ, then the reflection coefficient is $K = \exp(j\delta)$, and Equation (9.16) may be rewritten as

(9.19)





The standing waves have nodes at the positions xn where cos(kxn + δ) is zero. Restricting δ, to the range from −π to + π, and recalling that xn is a negative quantity since the line ends at x = 0, we find

(9.20)

If a line is terminated by an impedance which causes partial reflection, standing waves are superimposed on the part of the travelling wave which is absorbed. The amplitude of the alternating voltage across the conductors varies along the line in the way shown in Figure 9.8, with maxima and minima each having a spacing of half a wavelength. Writing the reflection coefficient as K = Kejδ, with K 1, Equation (9.16) can be split into its travelling wave and standing wave parts;

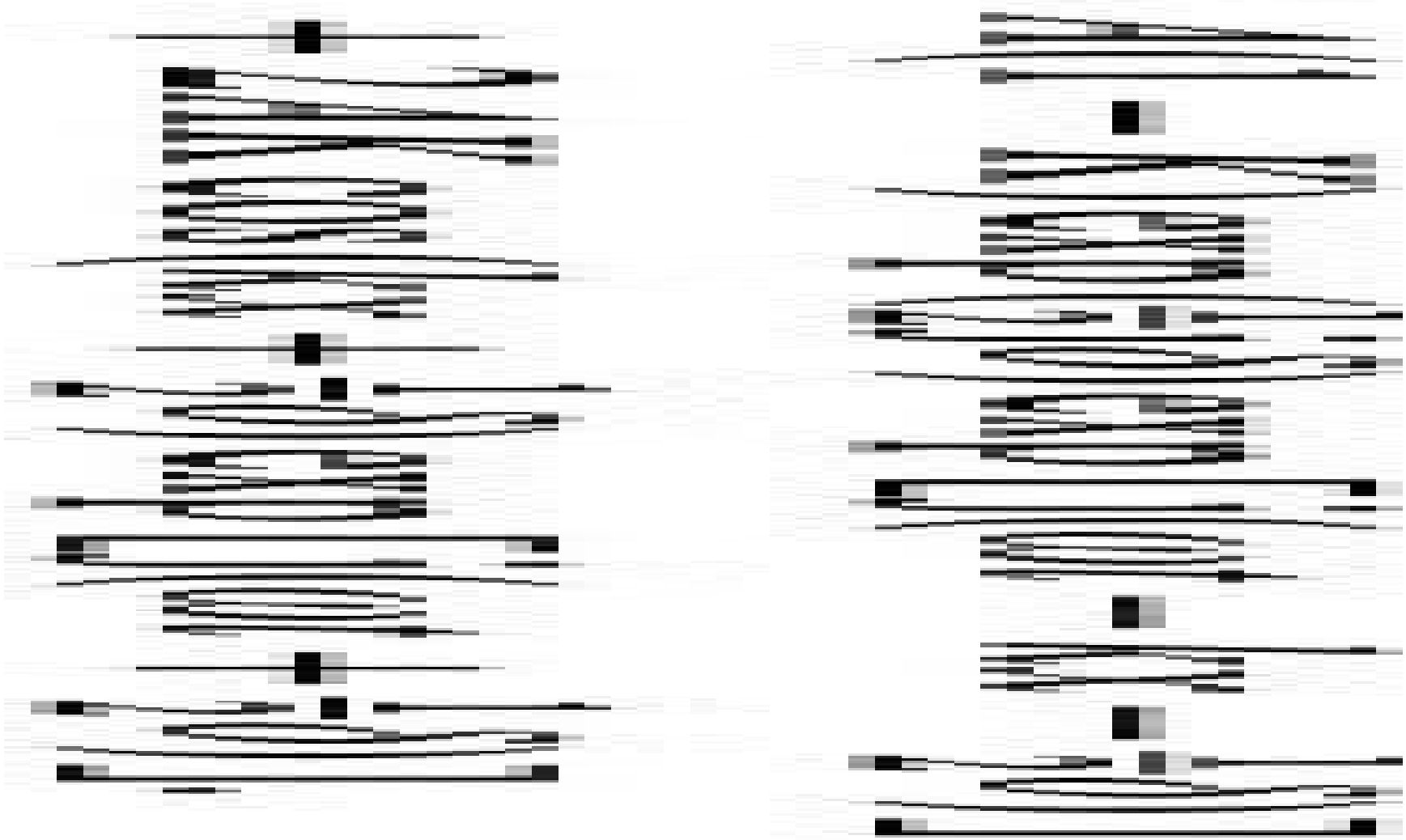
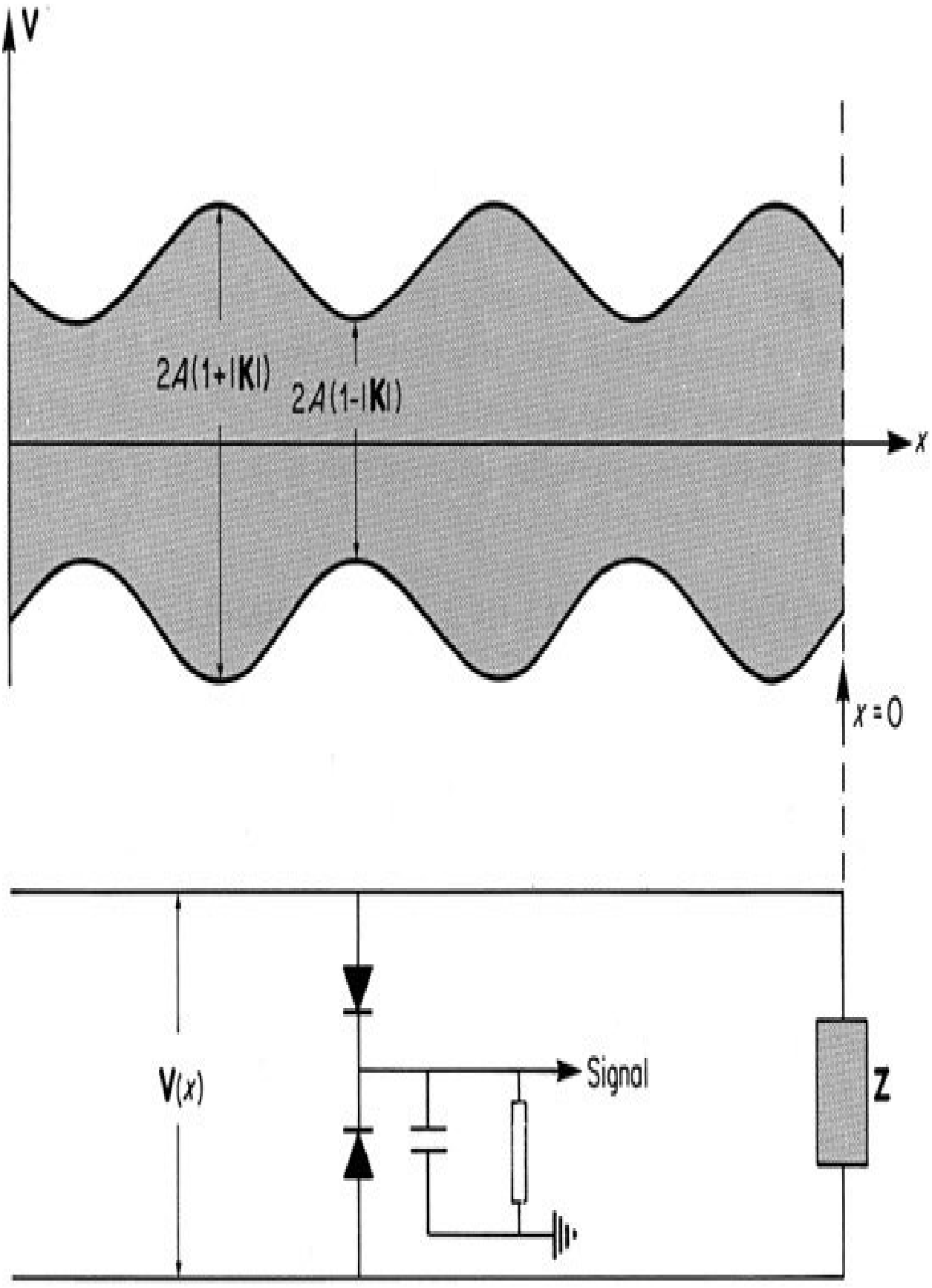
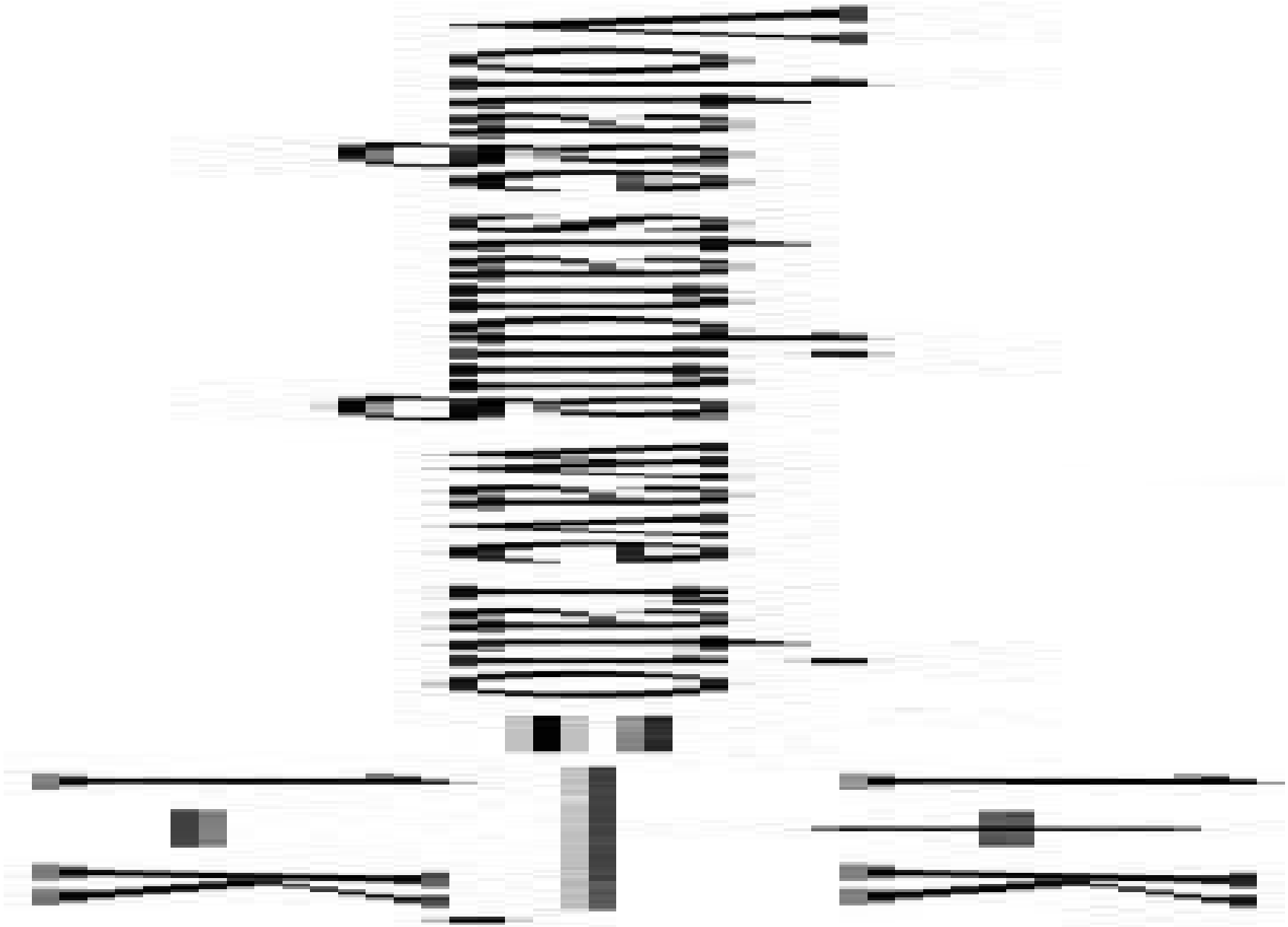


Figure 9.8. Measurement of the voltage standing-wave ratio on a mismatched transmission-line.


V
2A(1+|K|)
2A(1-|K|)
x
x = 0
V(x)
Signal
Z

The position of the minima is still related to the phase of the reflected wave by Equation (9.20), while the maximum amplitude of the voltage is A(1 + K) and the minimum is A(1 − K). Their ratio is called the voltage standing wave ratio:

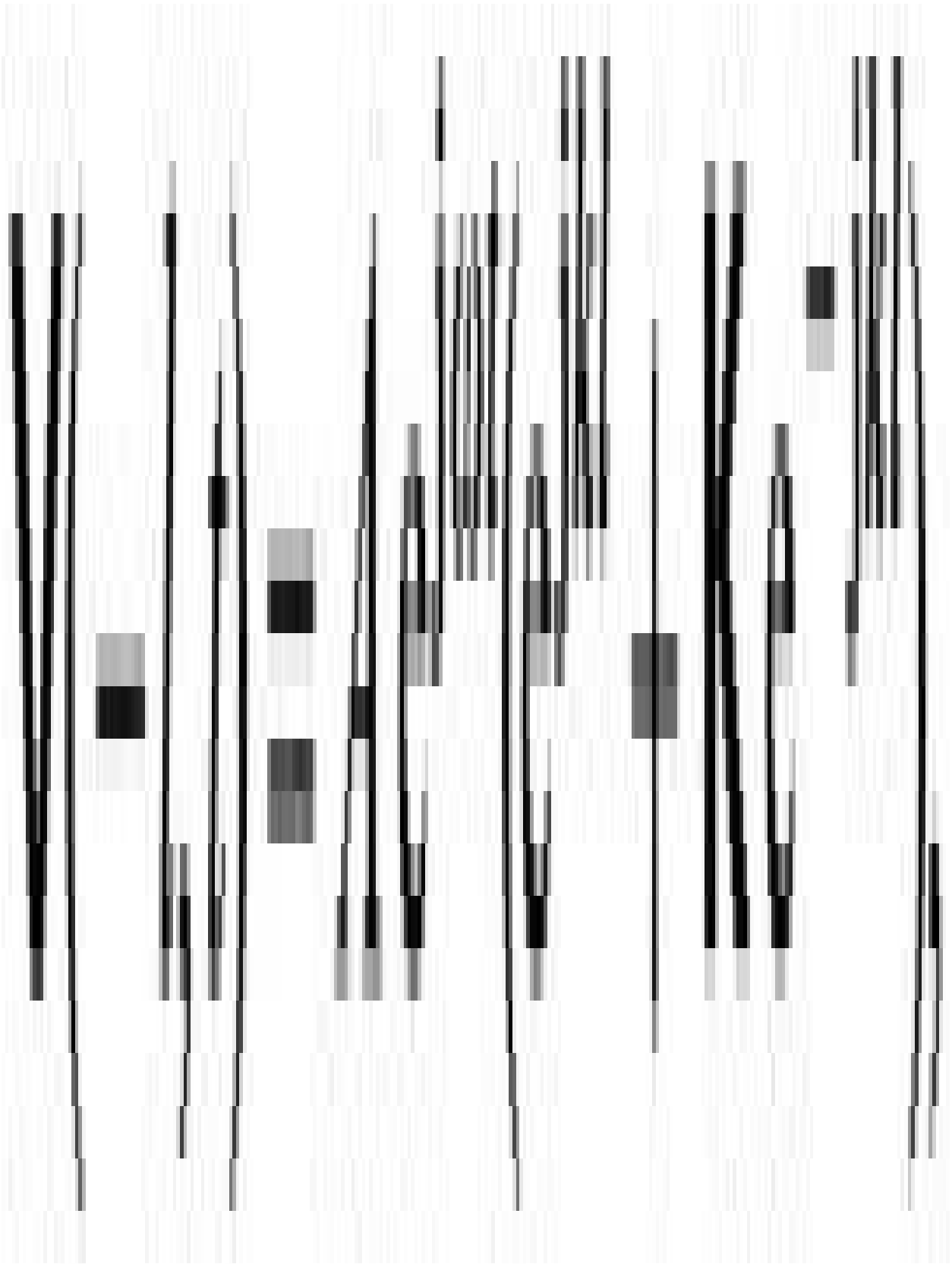
(9.21)



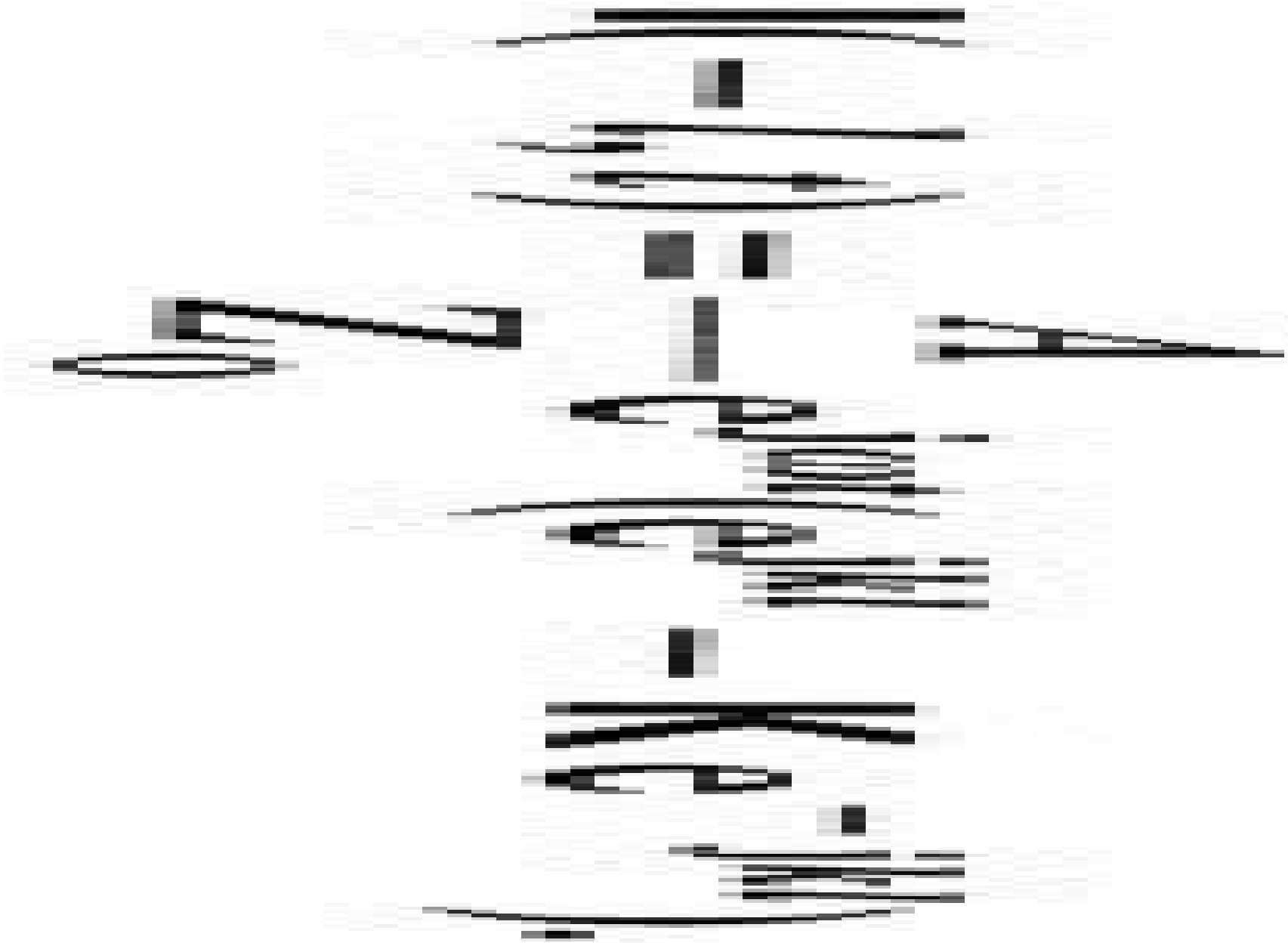
The voltages on a parallel wire transmission line can be detected by putting a rectifying circuit across the line as shown in Figure 9.8. The current drawn by this circuit must be very small so that the line voltage is not disturbed. The magnitude of the rectified current depends on the shape of the characteristic curve of the diodes, but provided the current is small enough, it can be shown to be proportional to the square of the voltage across the line. Both the positions of the voltage minima and the voltage standing wave ratio can therefore be found from measurements of the rectified current. The magnitude and phase of the reflection coefficient K are then known from Equations (9.20) and (9.21), and the value of the impedance terminating the line follows from Equation (9.18). This is a useful method of measuring impedances at frequencies of a few hundred MHz: at these very high frequencies A.C. bridges are not suitable, and the wavelength on the transmission line is conveniently of the order of one metre.

9.4 THE INPUT IMPEDANCE OF A MISMATCHED LINE

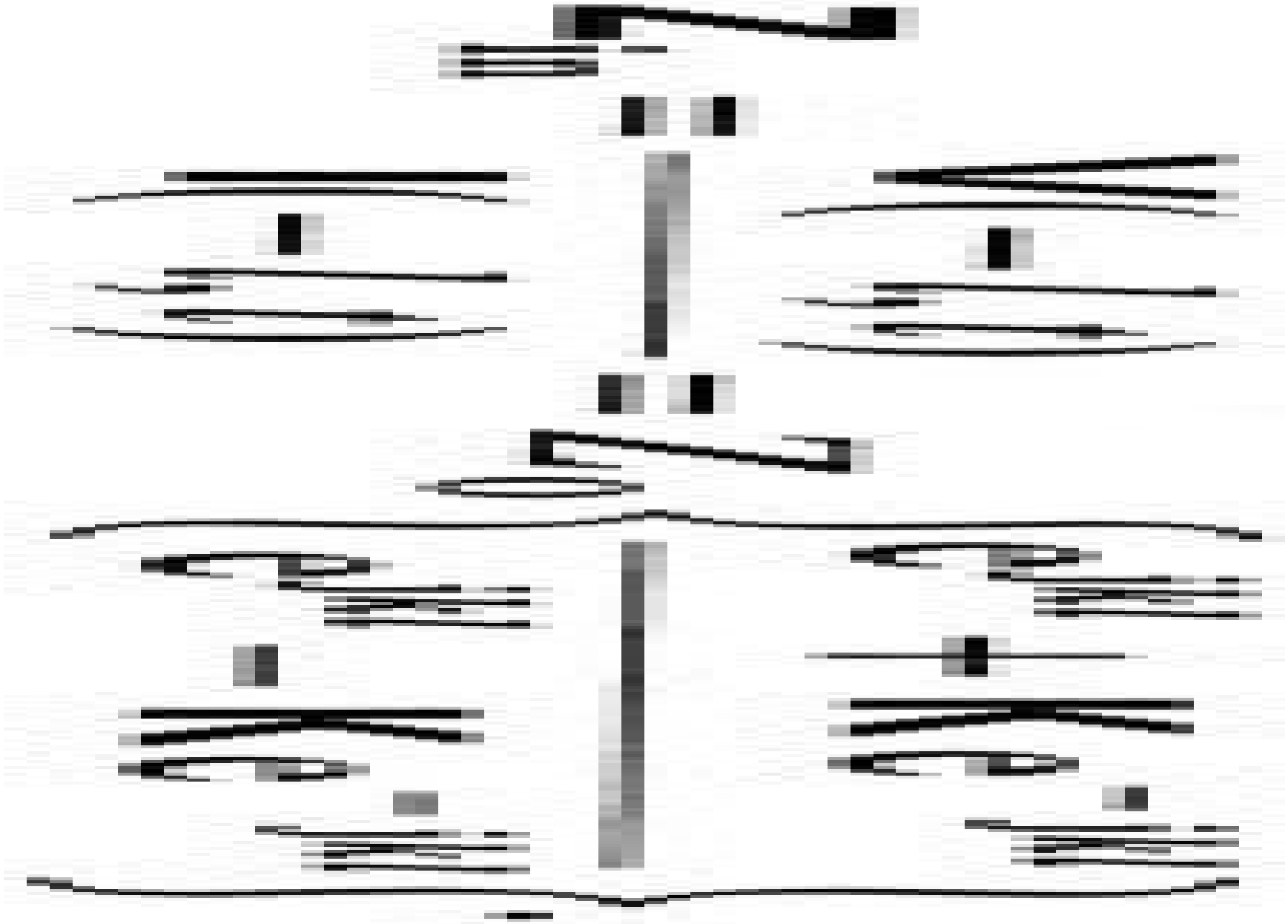
The input impedance of a mismatched line is found from the voltage and current at its input end. For a line of length l, the input is at $x = (-l)$, and from Equations (9.16) and (9.17) the voltage and current at this point are



and



The input impedance is



Substituting the value of K given by Equation (9.18), this expression becomes

(9.22)

Three cases are of special interest:

(i) Z = 0. For the shorted line equation (9.22) reduces to

(9.23)

The input impedance is purely reactive, because, as we have already pointed out above, no power can be dissipated at the shorted end of the line.

(ii) Z = ∞. Similarly the input impedance of an open-circuited line is purely reactive, and

(9.24)

The open-circuited line is thus equivalent to a shorted line differing in length by a quarter of a wavelength.

(iii) kl = π/2. Equation (9.22) now reduces to

(9.25)

The input impedance of the quarter-wave line is real when it is terminated by any resistor. It may appear to be paradoxical that the input impedance becomes very small when the line is terminated by a large resistance. For a fixed input voltage, according to Equation (9.25) the power delivered by the generator becomes indefinitely larger and larger as the terminating resistance R increases. However, we have not taken into account the output impedance of the generator. The voltage at the input of the line falls when the current demand becomes high and the voltage at the input end of an open-circuited quarter-wave lossless line must in fact always be zero, so that no power at all is delivered from the generator to the line.

Impedance matching

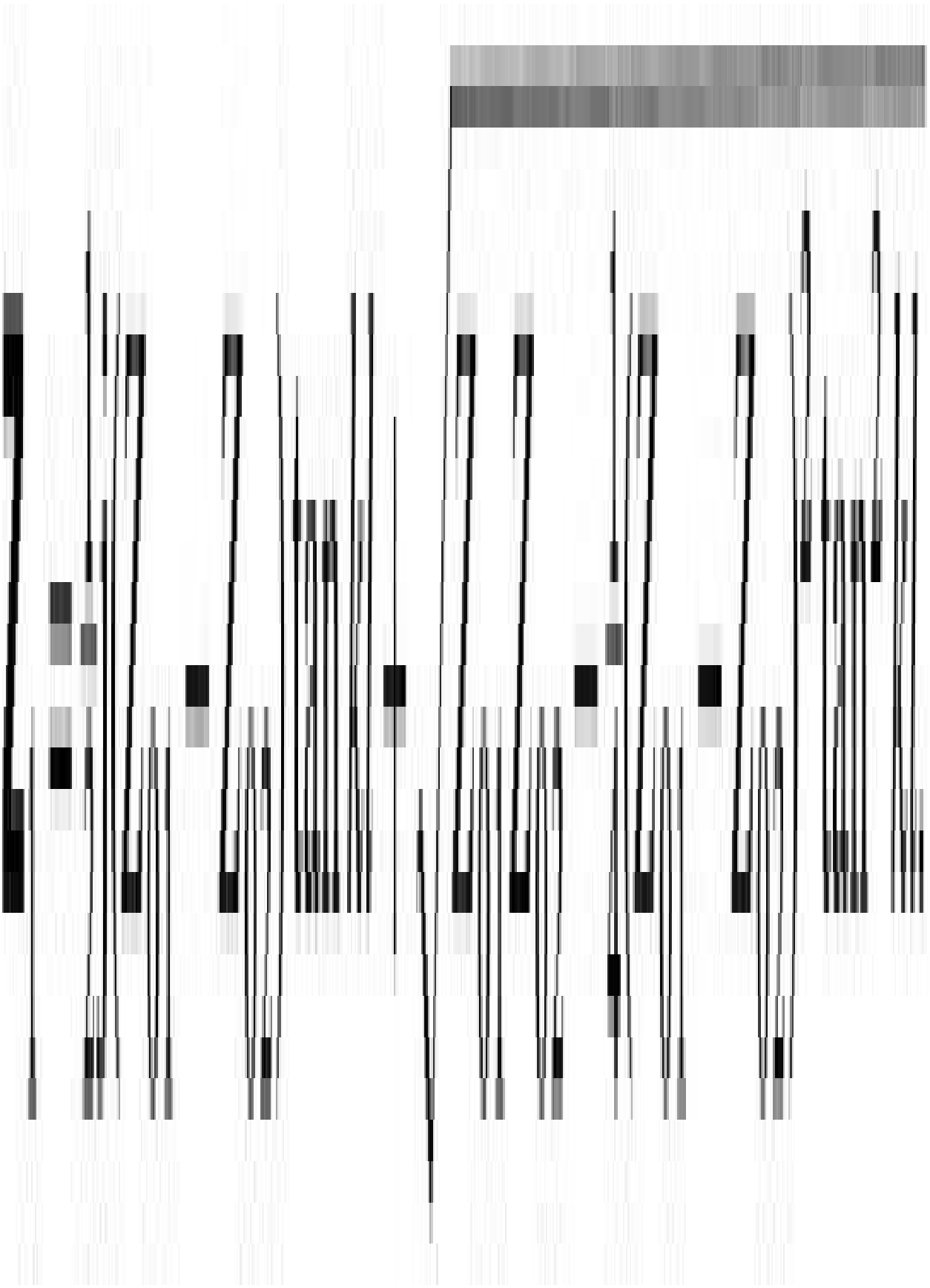
The power output of a generator depends on its output impedance, and it is easily shown that a generator delivers maximum power to a load which is equal to the complex conjugate of its own output impedance Zg. If Zg and the load R are both purely resistive and the value of is in the range of the characteristic impedances of practical transmission lines, then a quarter-wave line with Z0 = has an input impedance Zg when joined to the load, and may be used to maximize the power transfer from the generator.

The quarter-wave line is here acting as an impedance transformer. It can be used in a similar way to match a load to another line of different characteristic impedance. Suppose, for example, that a long transmission line of characteristic impedance is being used to transmit power to a load of resistance R≠ .

Reflections can be eliminated by inserting before the load a quarter-wavelength section of a line with characteristic impedance Z0 satisfying . The input impedance of the quarter-wave section is , i.e. the long line is matched and no reflected wave travels along it. Notice that the matching is only achieved at the particular frequency for which the additional section of line is exactly a quarter-wavelength long.

Transmission lines as high-frequency circuit elements

Generally there are reflections whenever transmission lines of different characteristic impedances are connected together, except in special cases like the example above where the two lines are matched at a single frequency. But it is possible to make a low-pass filter transmitting a wide range of frequencies by connecting together alternate sections of two different transmission lines. The result is a ladder rather like the ladder network made of lumped capacitors and inductors discussed in section 8.4.1, Figure 9.9 is a plan view of such a ladder constructed of strip lines mounted on a single slab of dielectric, with alternate sections of narrow and broad strips, all of equal length l. The narrow strips have a small capacitance per unit length, and hence a high characteristic impedance Z01, while the wide strips have a lower characteristic impedance Z02. Applying Equation (9.22) we can find the characteristic impedance of the whole ladder by requiring that the input impedance remains unchanged if another pair of narrow and broad strips is added to an existing matched ladder. It turns out that the input impedance of a ladder starting with a narrow section is



This impedance becomes purely imaginary if

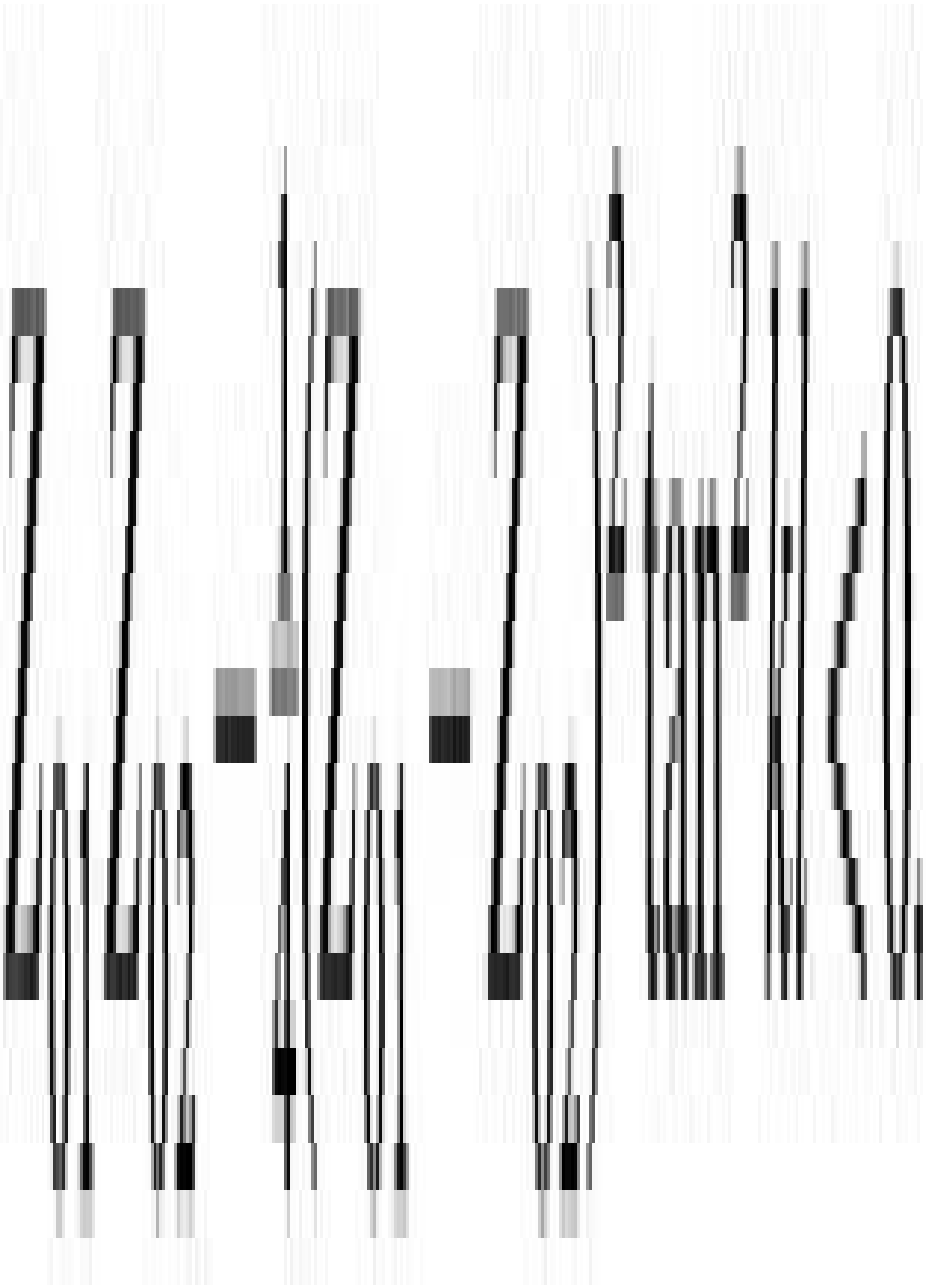
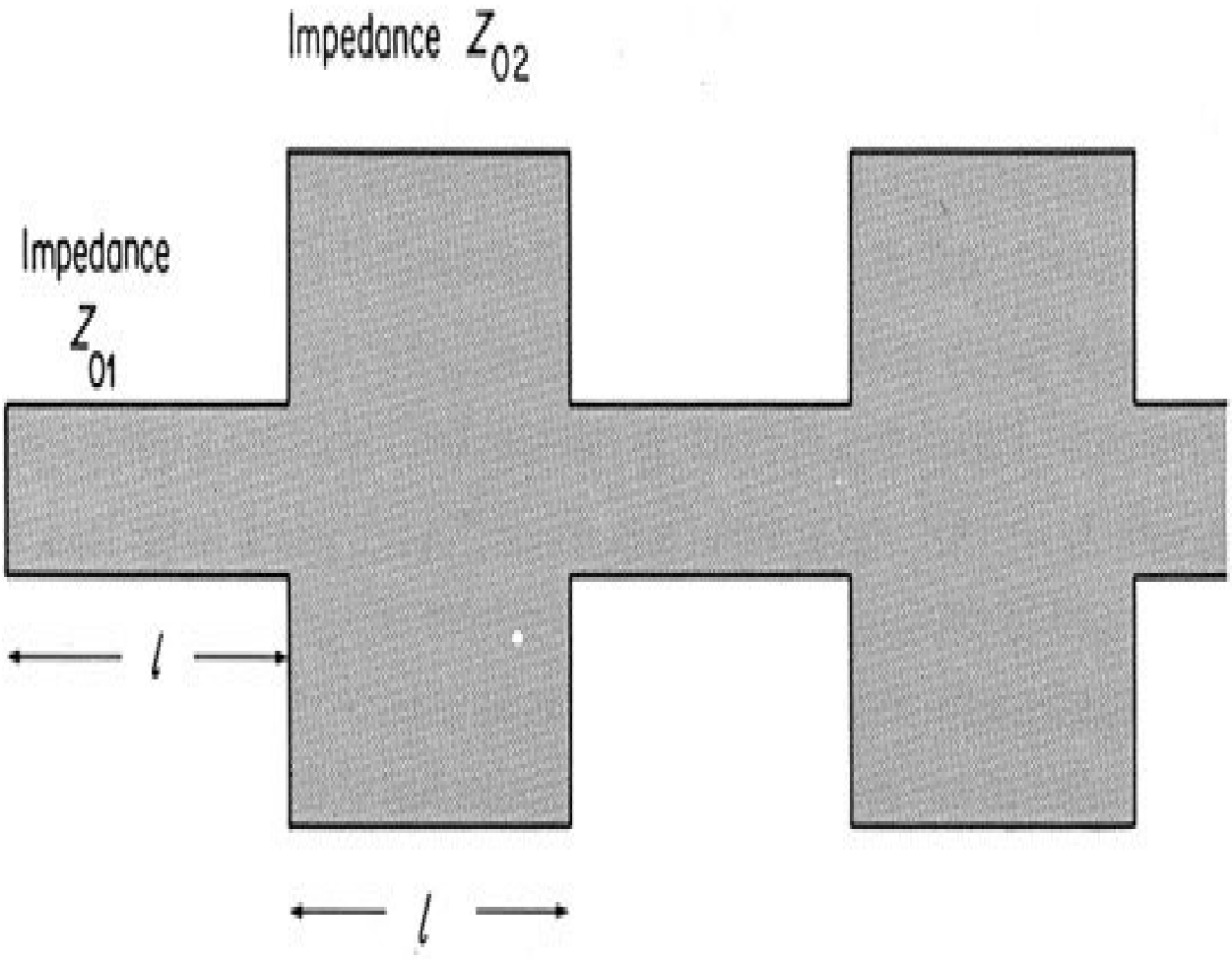
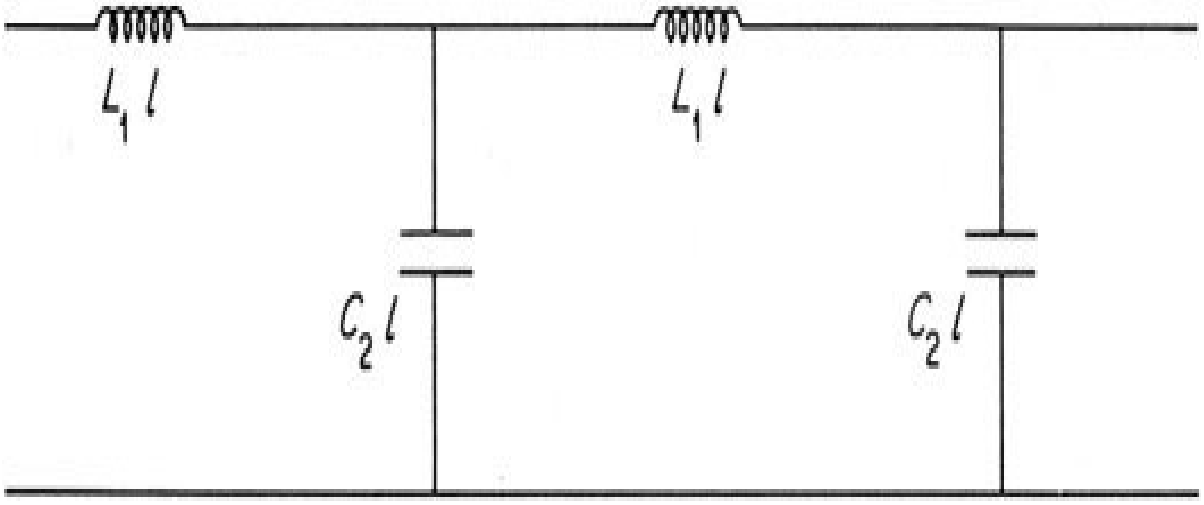


Figure 9.9. A strip line filter can be made with conductors consisting of thin sheets of the shape shown by the shaded area: the conductors are separated by a dielectric slab of constant thickness. Signals travelling from left to right meet alternate narrow and broad sections with characteristic impedances Z01 and Z02. The approximate equivalent circuit represents the narrow sections as idealized inductances and the broad sections as lumped capacitors.


Impedance Z_{02}
Impedance
Z_{01}
l
l


$L_1 l$
$L_1 l$
$C_2 l$
$C_2 l$

When this inequality is satisfied there is total reflection, and an input signal is attenuated as it passes along the ladder. On the other hand, when the quantity under the square root sign is positive, the input impedance is partly resistive, implying that as for the LC ladder network, the input signal is transmitted without attenuation. There is a series of pass-bands because of the cyclic behaviour of tan kl, and there are cut-off angular frequencies ωc satisfying the condition

$$\tan^2(k_c l) = \tan^2\left(\frac{\omega_c l\sqrt{\varepsilon}}{c}\right)$$

$$= \frac{4Z_{01}Z_{02}}{(Z_{01}-Z_{02})^2}$$

where ε is the relative permittivity of the insulator, and the propagation velocity on the line.

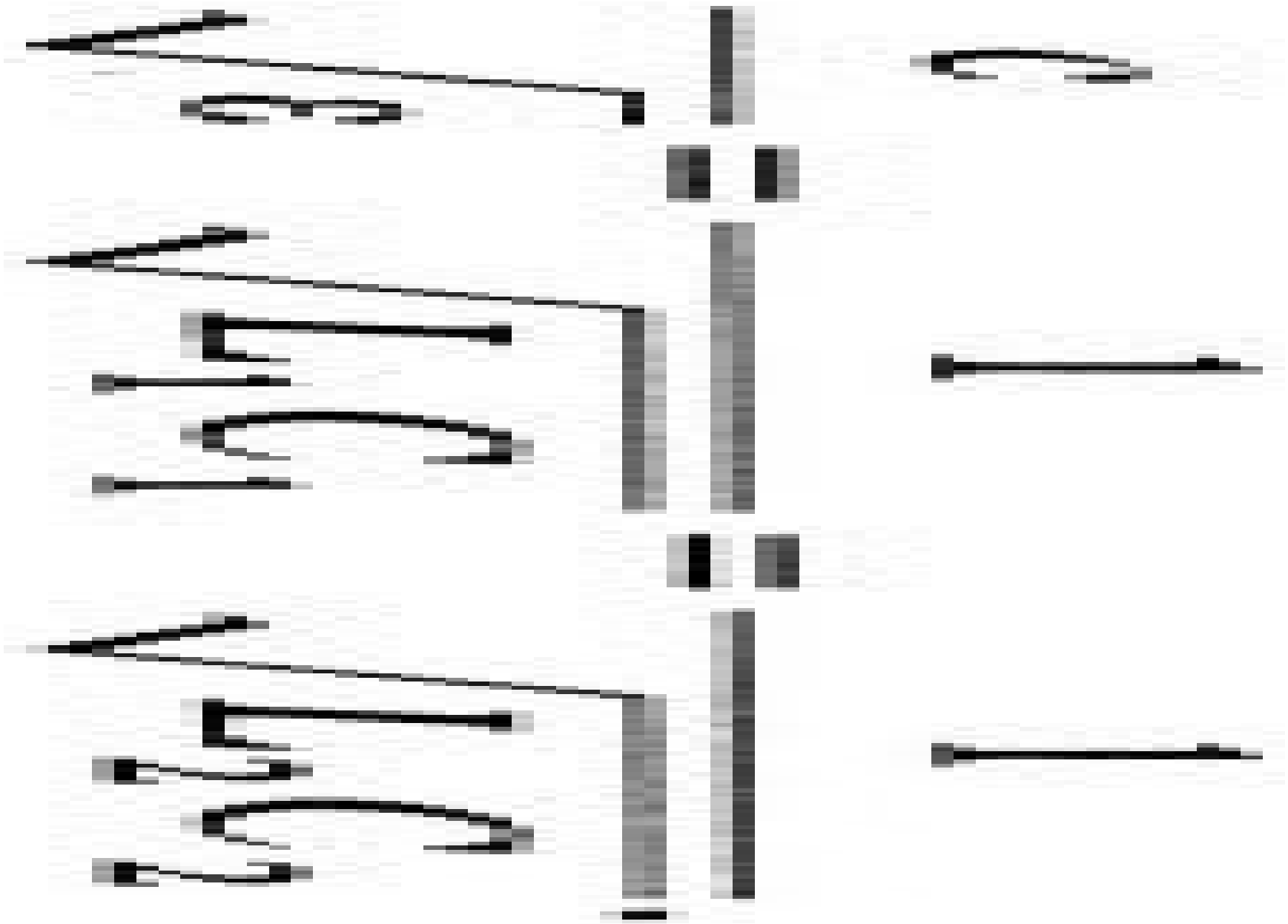
Let us concentrate on the lowest pass-band, for which $kl < \pi/2$, and assume that the wide strips are much wider than the narrow strips. Because the characteristic impedance is proportional to the strip width, we can put (Z01 − Z02) Z01, and the cut-off condition becomes

$$\tan^{-1}\left(\frac{\omega_e v_e}{c}\sqrt{\frac{4z_{02}}{z_{01}}-1}\right)$$

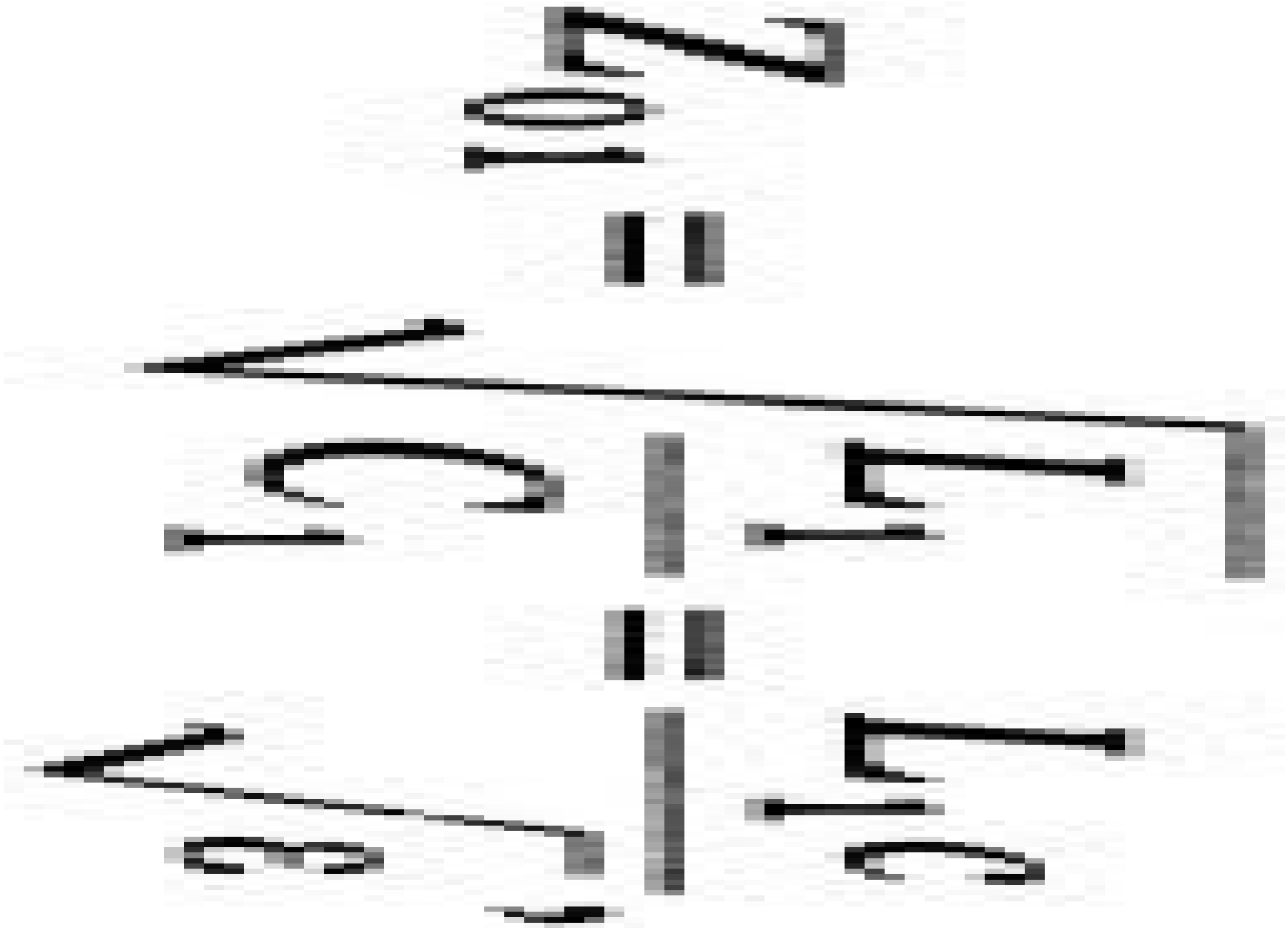
or, making the approximation tan θ θ,

(9.26)

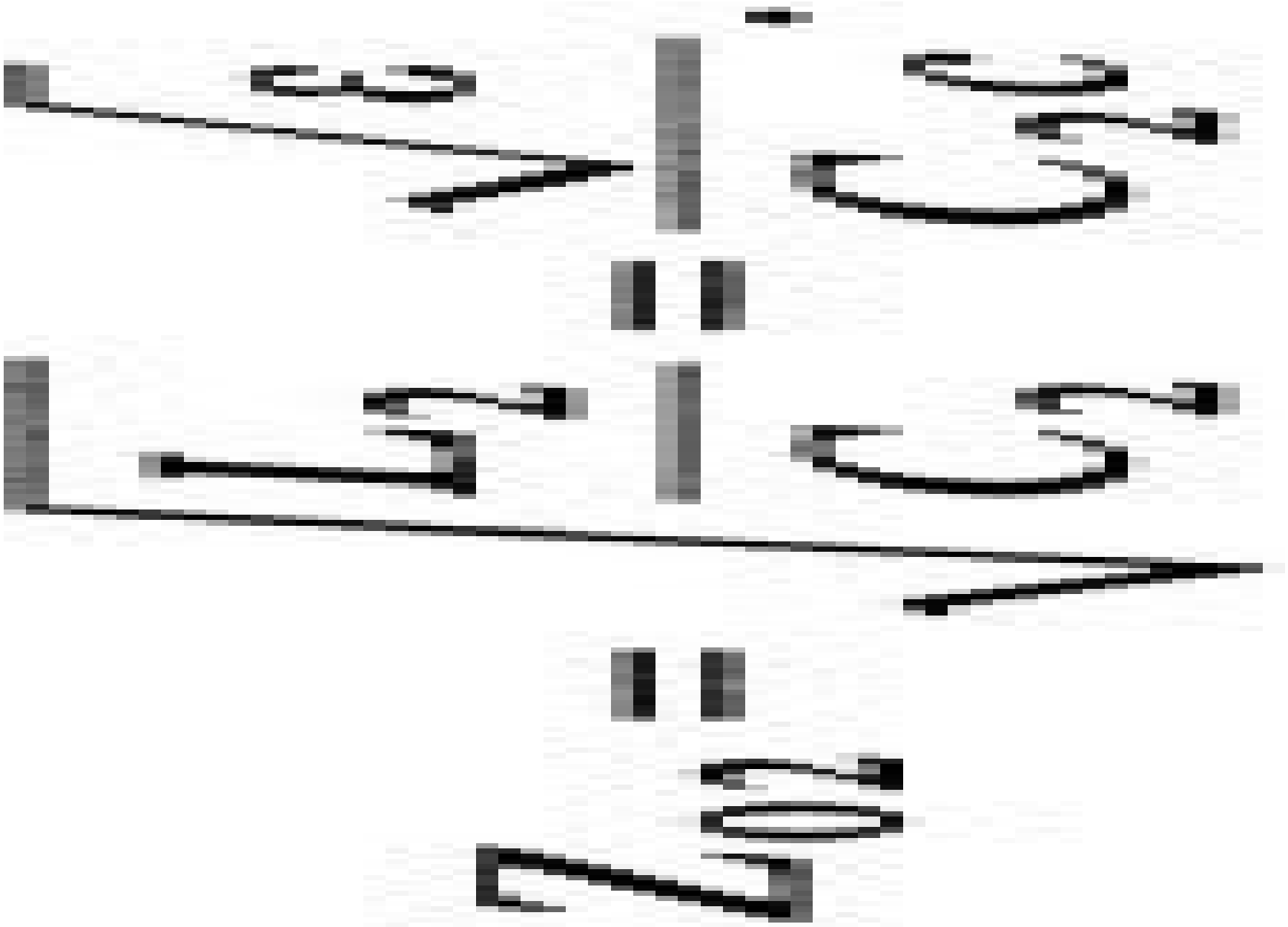
For example, if Z01 = 4Z02, the length of each section of the filter is 5cm and the relative permittivity ε = 2.3, the cut-off frequency ωc/2π is roughly 400 MHz. Although this frequency is so high, the transmission line filter really works in a very similar way to the LC filter discussed in section 8.4. A section of the wide strip has a much higher shunt capacitance than a section of the narrow strip, and conversely, referring back to Equation (9.14), one sees that the narrow strip has the larger series self-inductance. The sections of transmission line can be pictured as the components of a network, the wide strips representing capacitors and the narrow ones inductances, as shown in Figure 9.9. In the spirit of this picture, let us rewrite the cut-off condition (9.26) in terms of the capacitance per unit length C2 of the wide strip and the self-inductances per unit length L1 of the narrow strips. Signals travel along both strips at the same speed , given in terms of the line parameters by the equations



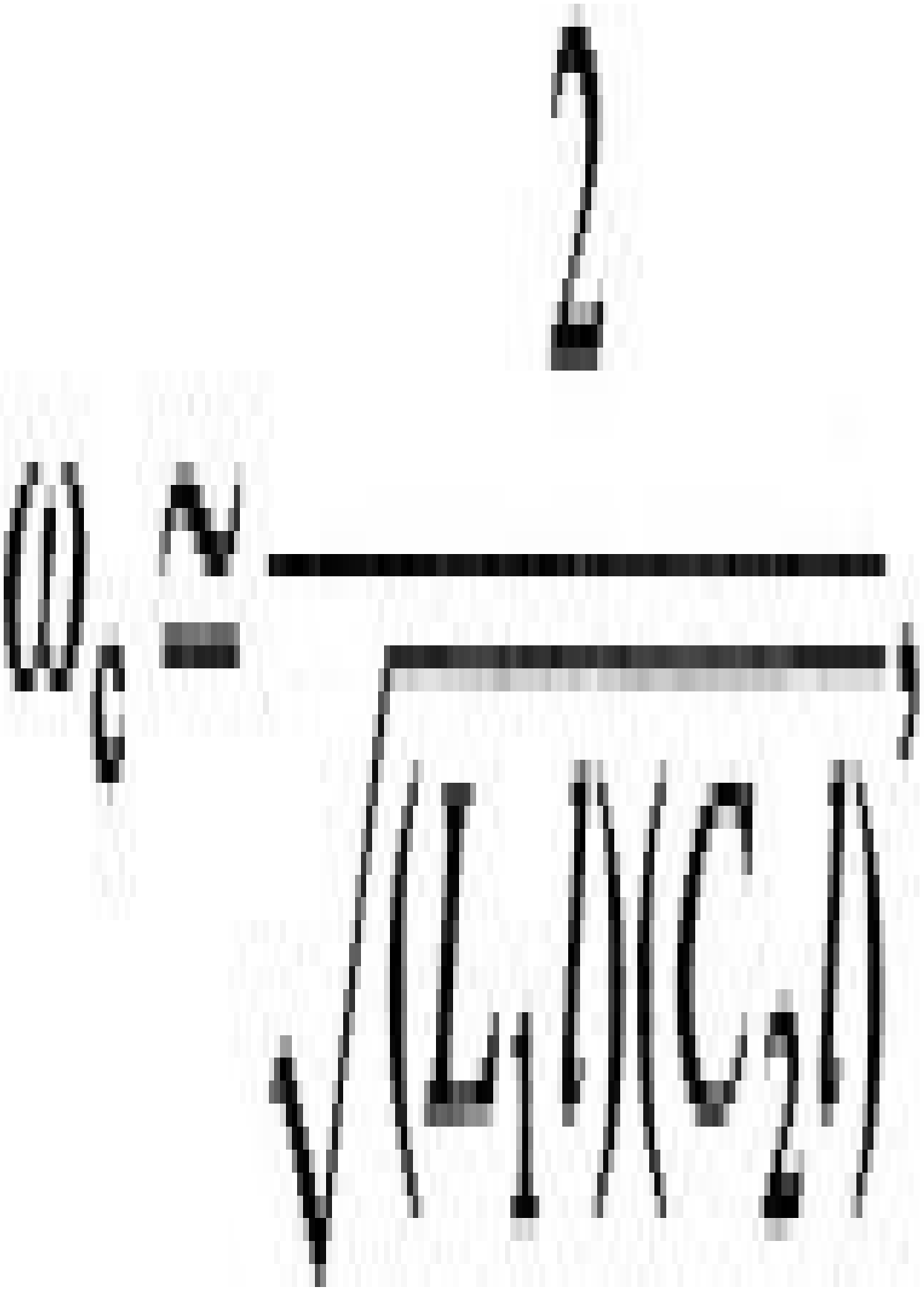
Using these equations to eliminate C1 and L2, we may put



and



The cut-off condition Equation (9.26) becomes



exactly the condition obtained in section 8.4 for the lumped parameter filter, since (L1l) and (C2l) are respectively the total inductance and total capacitance of the neighbouring sections of transmission line. This is a good example of how at high frequencies the role of the components used at low frequencies can be taken over by lengths of transmission line.

9.5 LOSSY LINES

For most purposes the assumption that a transmission line is lossless is a very good approximation. At a frequency of 1 MHz an alternating voltage applied to the input of a typical laboratory coaxial cable is attenuated by considerably less than 0.1% per metre of cable. Attenuation can safely be ignored when signals at about this frequency are transmitted from one piece of apparatus to another along a few metres of cable. However, attenuation may become important in the following circumstances:

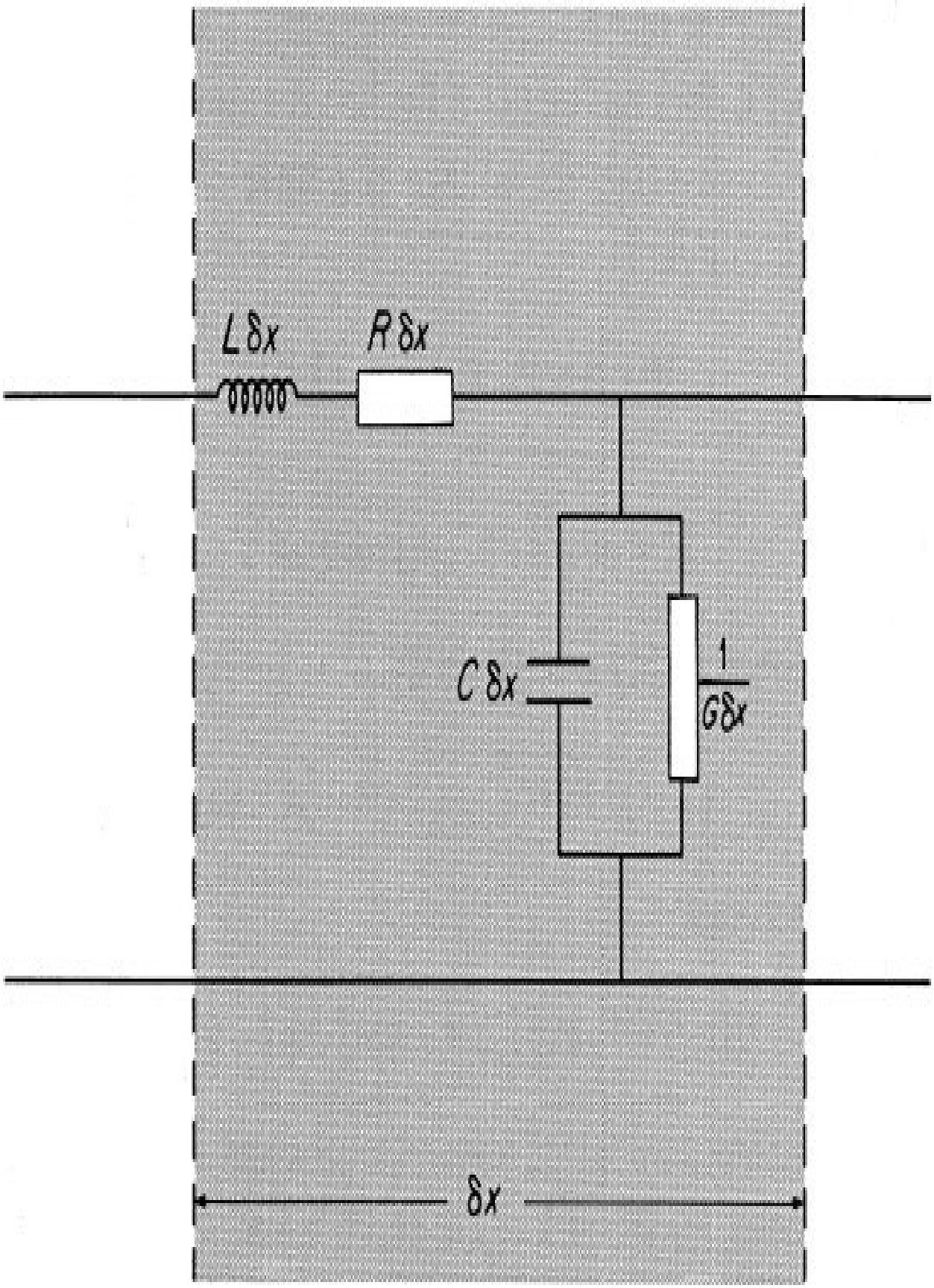
(i) when signals are to be transmitted for great distances along the cable.

(ii) at very high frequencies, since it turns out that the losses increase with frequency.

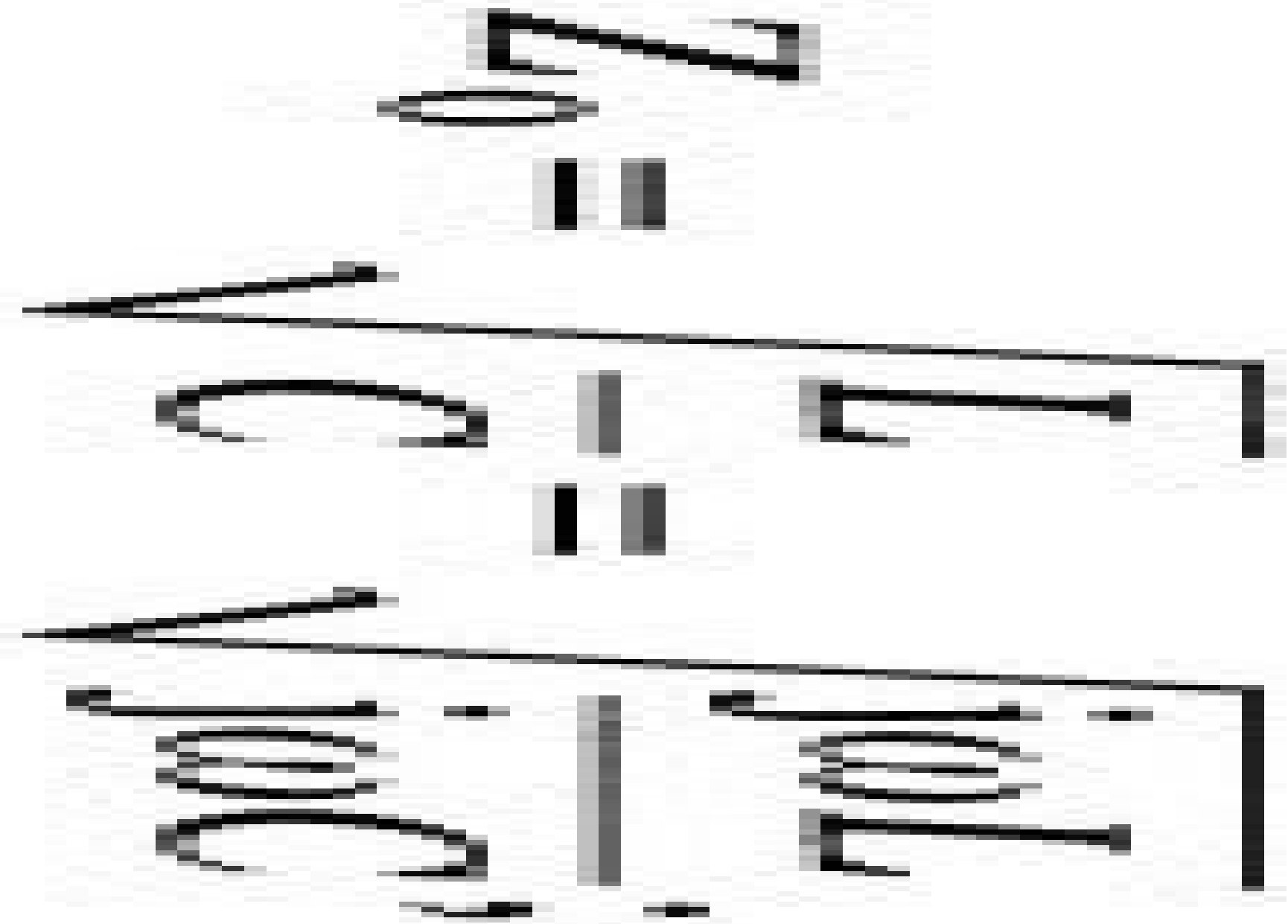
(iii) when the line is handling large amounts of power. The power lost in transmission is dissipated as heat, and cooling the line may be a problem.

Attenuation arises because of the non-zero resistance of the conductor and the finite resistance of the insulator from which the transmission line is constructed. As we did when setting up the differential equations for the lossless line, let us consider the series and parallel impedances of a section of line of length δx, now including the effects of resistance. As before we write the self-inductance per unit length as L and the capacitance per unit length as C. If the resistance per unit length of the conductor is R, a resistance $R\delta x$ must be added to the self-inductance $L\delta x$, and the total series impedance to a signal of angular frequency ω is $(R + j\omega L)\delta x$. Similarly if the conductance per unit length of the insulator is G (conductance is the inverse of resistance), a resistance $1/(G\delta x)$ is added in parallel to the capacitance $C\delta x$. The admittance between the two conductors of the section δx is $(G + j\omega C)\delta x$. The equivalent circuit of the section δx is shown in Figure 9.10.

Figure 9.10. The equivalent circuit for a short length of lossy transmission line.


$L\delta x$
$R\delta x$
$C\delta x$
$\frac{1}{G\delta x}$
δx

The analysis of the propagation of signals along the line is exactly the same as for the lossless line except that the impedance $j\omega L$ is replaced by $(R + j\omega L)$, and the admittance $j\omega C$ is replaced by $(G + j\omega C)$. Now the characteristic impedance of a lossless line is

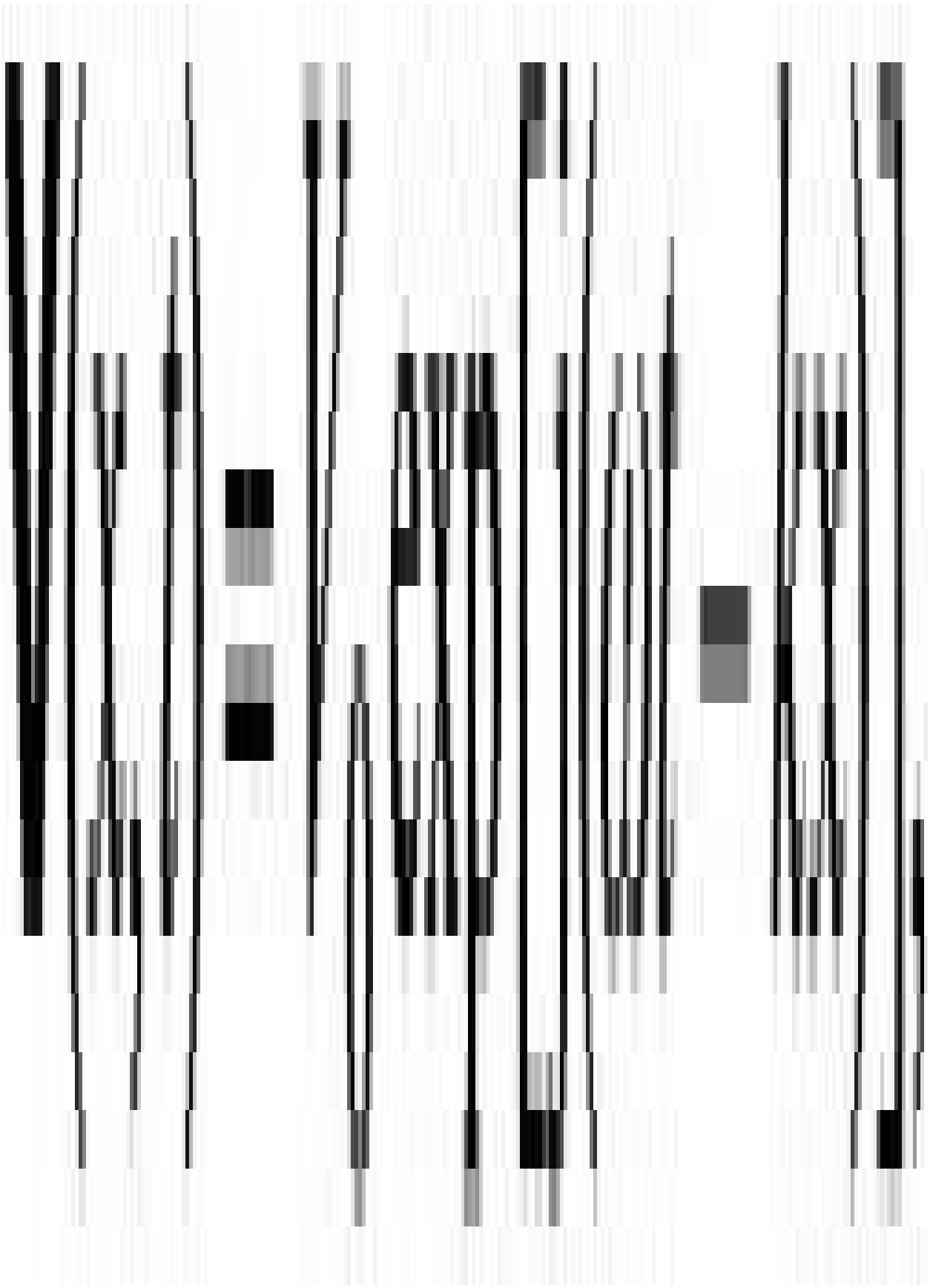


in a lossy line the characteristic impedance becomes complex and we write

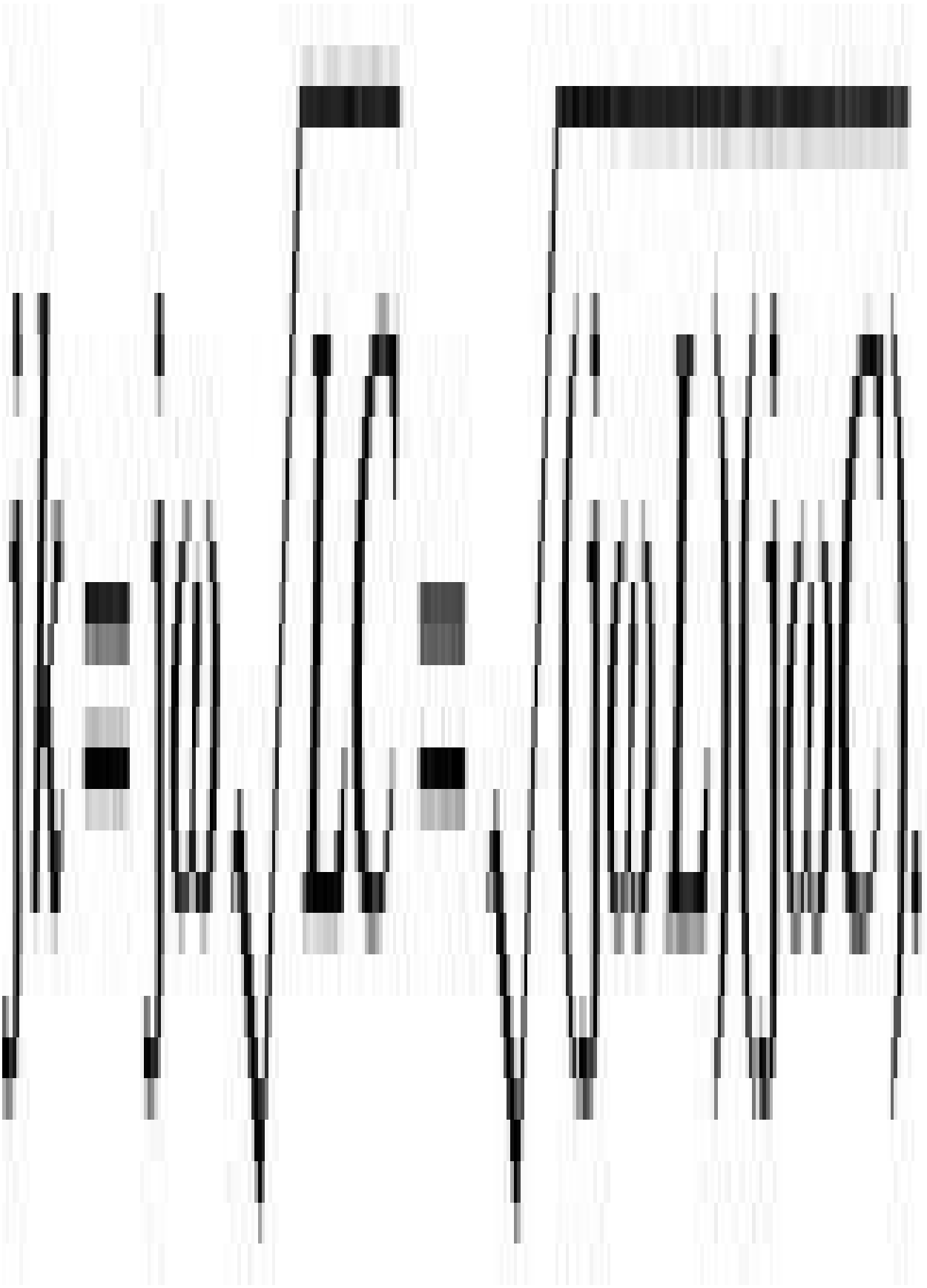
(9.27)

The characteristic impedance approaches its lossless value at high frequencies, although it is at high frequencies that the losses cause most attenuation. Usually reflections are not important even on a lossy line if it is terminated by an impedance : typically the imaginary part of the characteristic impedance is less than 10% of the real part at 1 MHz, and is decreasing with increasing frequency. For frequencies below 1 MHz the wavelength of the signals on the transmission line are so long that unless the line is very long the phase changes introduced by the line are negligible. Ordinary A.C. theory in which the line is regarded simply as a lead carrying currents is then adequate, and it does not matter if the line is mismatched.

For a matched lossless line, the voltage of a wave of angular frequency ω travelling from left to right is

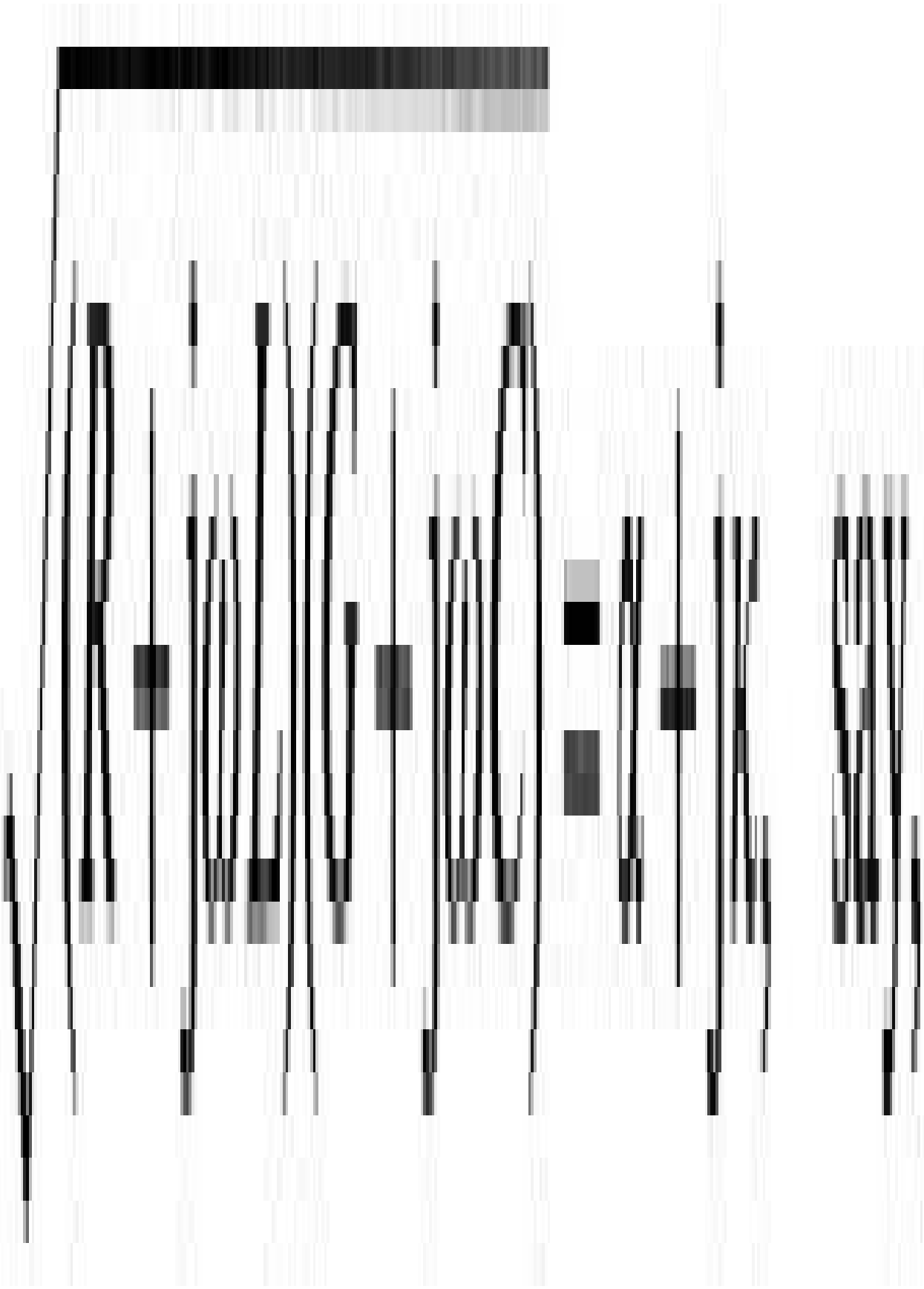


where



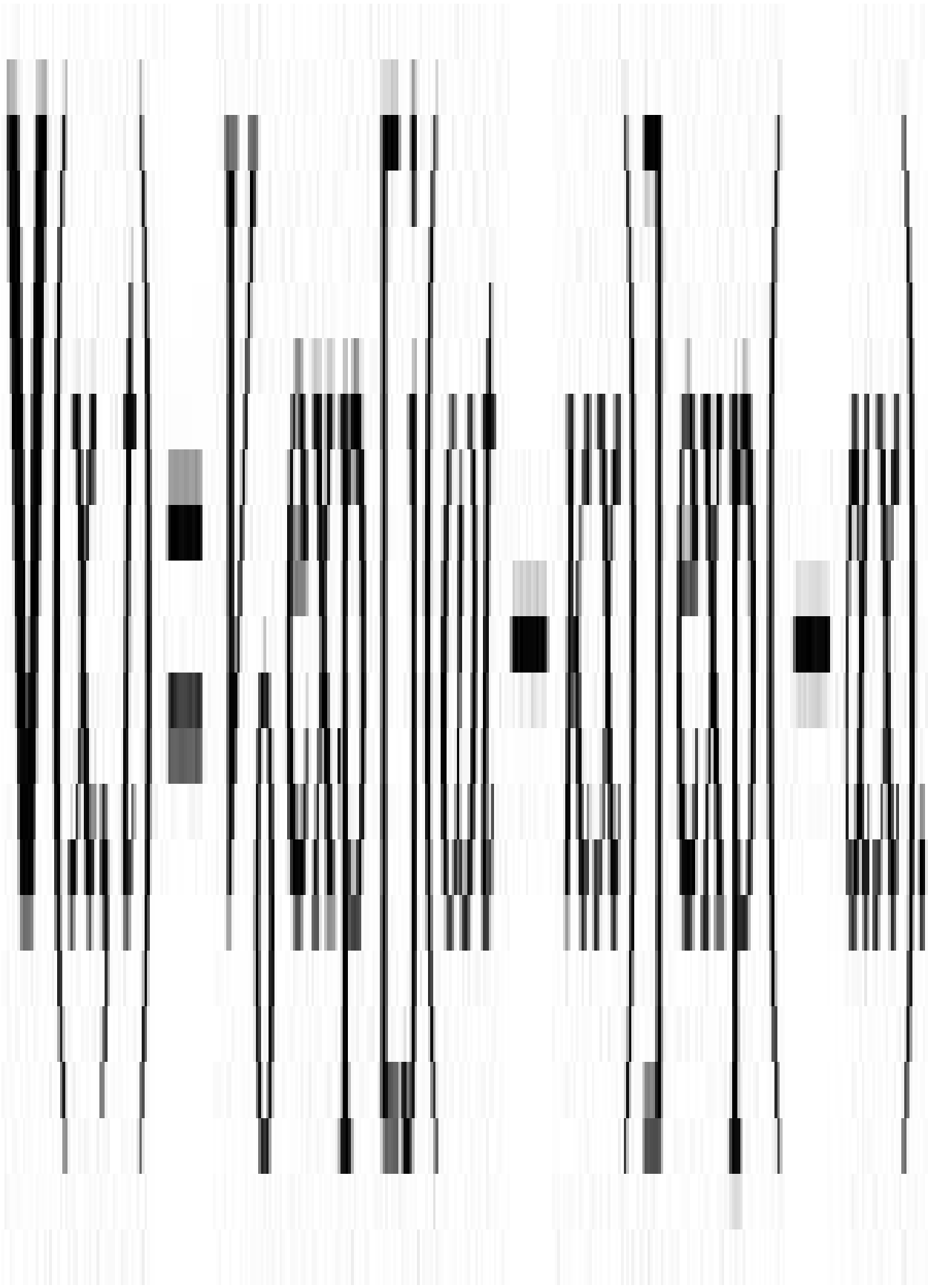
In the lossy line this quantity is replaced by

(9.28)



and the voltage becomes

(9.29)



This represents a damped wave with wave-number k and attenuation constant α.

At high frequencies, for which R « jωL and G « jωL, Equation (9.28) can be expanded to first order to give

(9.30)

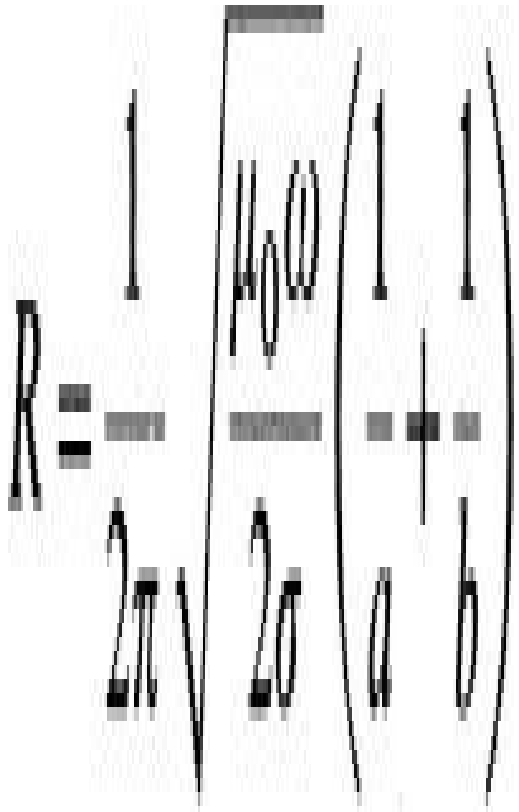
$$\alpha + j\kappa \approx j\omega\sqrt{LC} + \omega\sqrt{LC}\left\{\frac{R}{2\omega L} + \frac{G}{2\omega C}\right\}$$

$$\approx j\omega\sqrt{LC} + \frac{R}{2Z_0} + \frac{1}{2}GZ_0.$$

since is the high frequency value of Z0. Equation (9.30) shows that to first order the wave number k = ω , the same value as for the lossless line. Hence all frequencies propagate at the same speed 1/ . For a 1 metre length of a typical coaxial cable, the resistance between the conductors was calculated in section 4.1.2 to be about $3 \times 10^{11}\Omega$. The series resistance is mostly associated with the inner conductor, and for a 1 mm diameter copper wire is about $0.1\Omega m^{-1}$. Using these values for R and G in a coaxial cable, Equation (9.30) gives an attenuation length of 1 km for a characteristic impedance of 50Ω. This is too optimistic, because R and G both increase with frequency. There are losses in the dielectric at high frequency because of a hysteresis effect in materials with a permanent dipole moment. In the frequency range of interest for transmission lines, a roughly constant amount of work is done in taking the material once round the hysteresis loop, and the dielectric losses are therefore proportional to frequency. We shall not discuss dielectric losses further, because the attenuation of a transmission line is usually dominated by conductor losses.

The effective resistance per unit length of the conductors increases because of the skin effect which causes high frequency currents to flow in a shallow layer on the surface of the conductor. The skin effect is discussed in Chapter 11, where it is shown that at an angular frequency ω the current in a material of conductivity σ is effectively carried in a surface layer of thickness . The high frequency resistance per unit length of a coaxial cable, for example, is roughly the same as the D.C. resistance of unit length of two cylinders of thickness δ and radii a and b equal to the inner and outer radii of the cable, i.e.

(9.29)



$$\approx 1.7 \times 10^{-8} \sqrt{\omega \left(\frac{1}{a} + \frac{1}{b} \right)} \text{ ohms per metre for copper.}$$

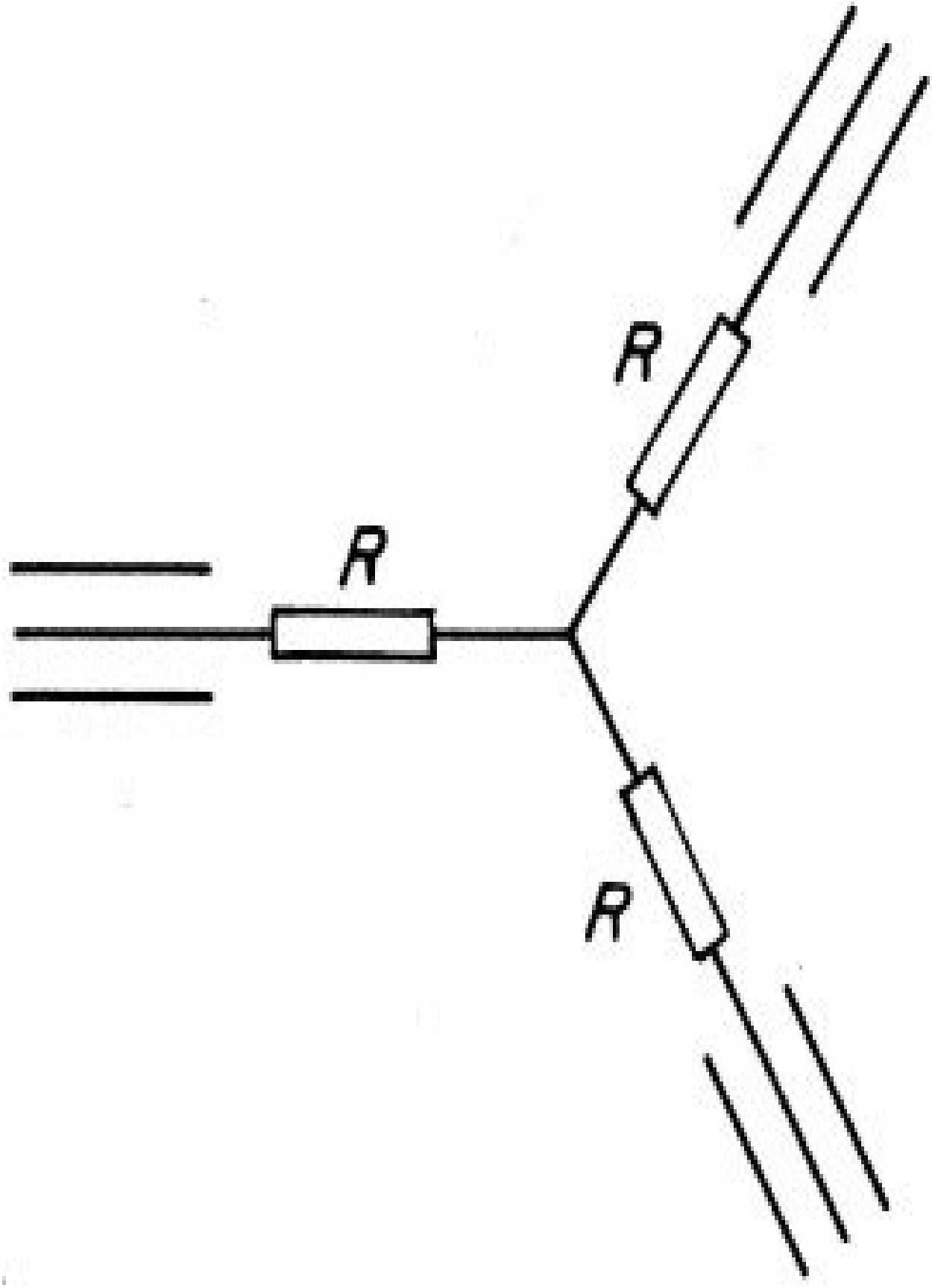
Most of the resistance is in the inner conductor, and if its diameter is 1 mm, . For a cable with characteristic impedance 50Ω, the attenuation length , or about 40 metres at a frequency of 1000 MHz.

PROBLEMS 9

9.1 A 50 MHz signal propagates along a transmission line which has a characteristic impedance of 200Ω. The voltage across the line is zero at a point 75cm from the end. What is the impedance terminating the line?

9.2 The central conductors of three 50Ω coaxial cables are joined together as shown in Figure 9.11. What must be the value of R if signals passing along any of the cables are to suffer no reflections? (The outer conductors are all earthed.)

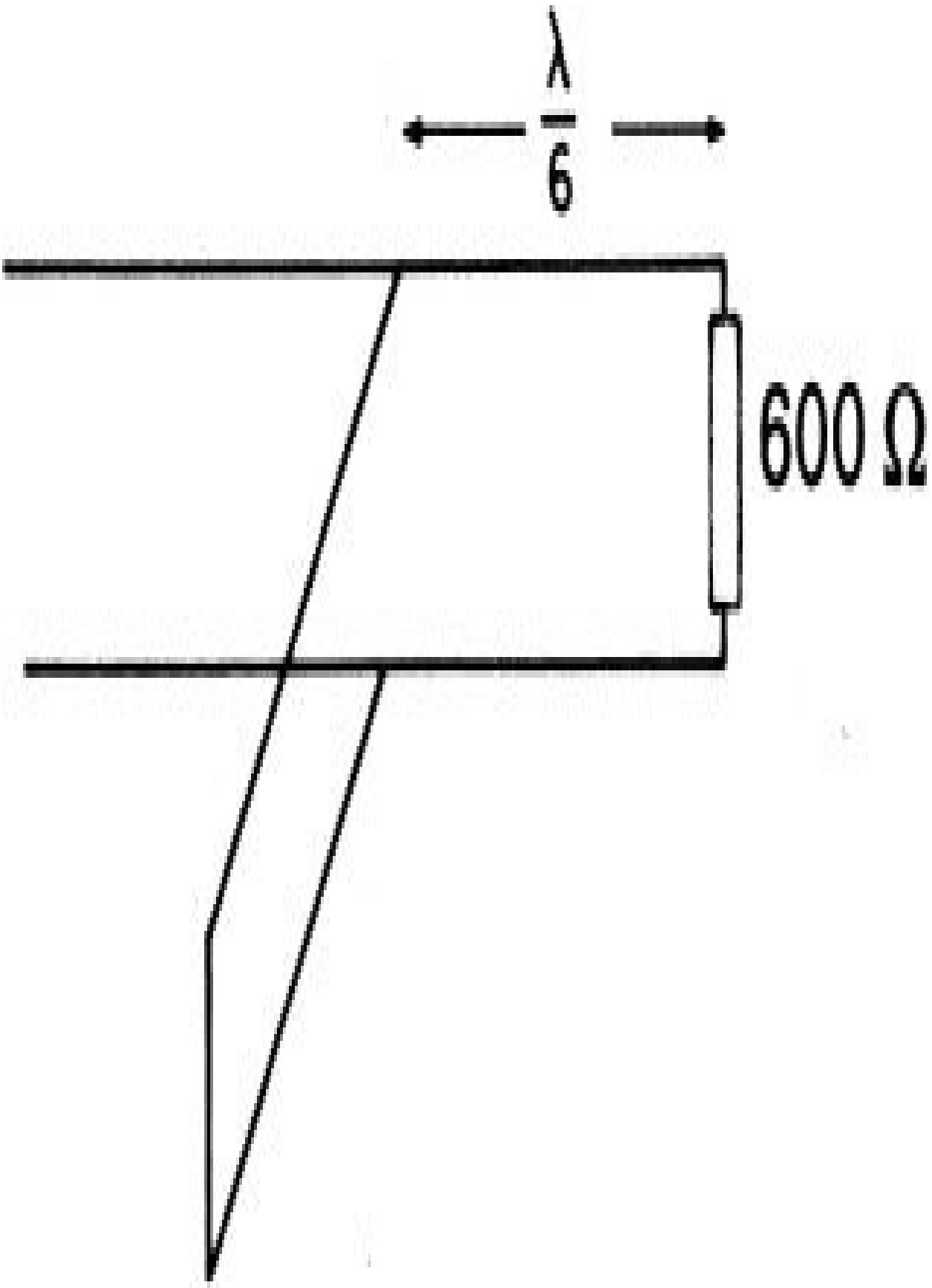
Figure 9.11.


R
R
R

9.3 Equal capacitors join the conductors at each end of a piece of 50Ω coaxial cable which is a quarter wavelength long at 10 MHz. If a source of e.m.f. of this frequency is applied across one of the capacitors, for what value of capacitance will the current drawn from the source be a minimum?

9.4 A 200Ω transmission line is terminated by a 600Ω resistor. What is the input admittance of the line at one-sixth of a wave-length from the end? A short-circuited stub of line is added in parallel at this point, as shown in Figure 9.12. What should be its length in order to minimize the reflected signal travelling back towards the source?

Figure 9.12.

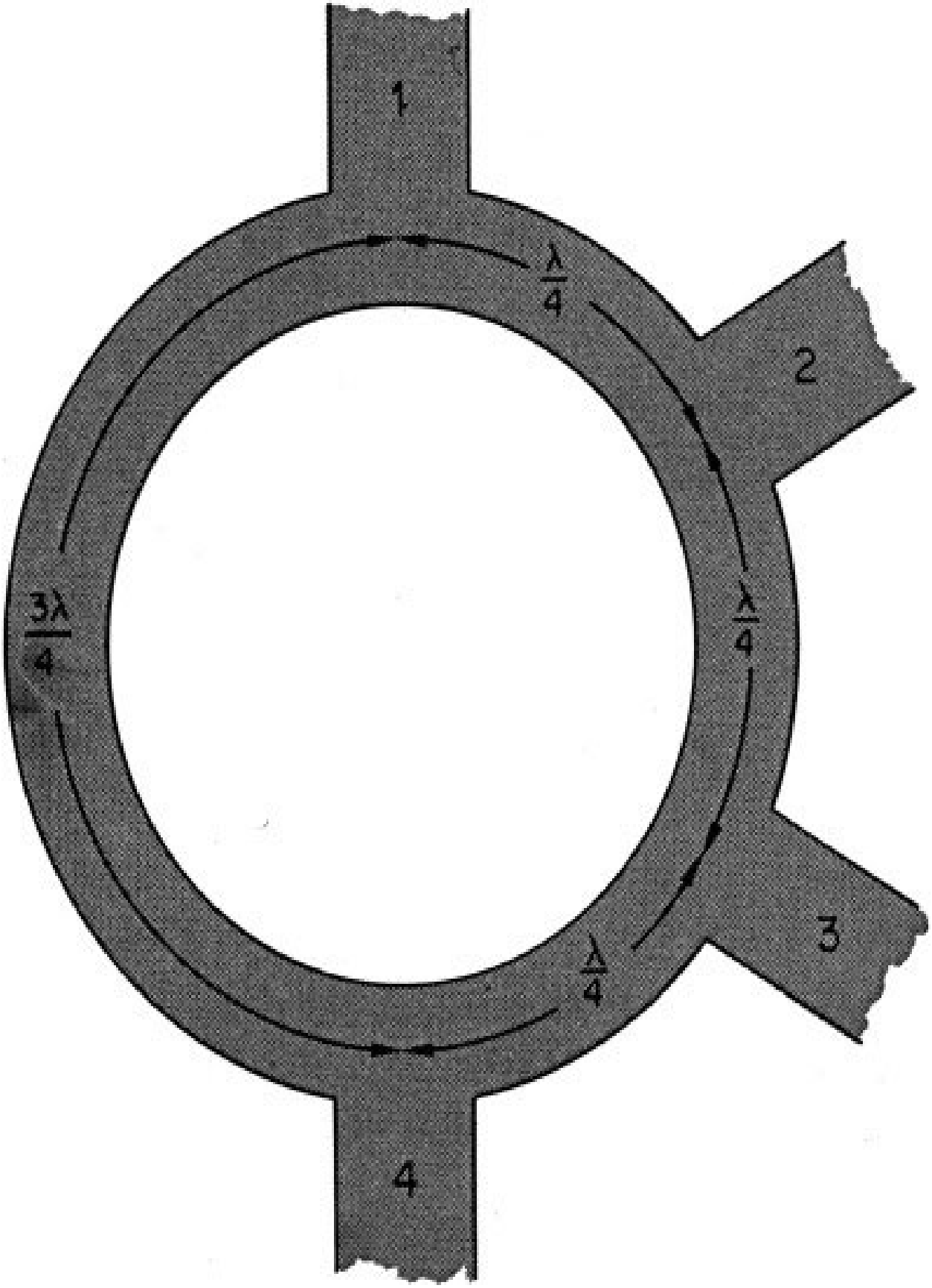

λ
6
600 Ω

9.5 A 50Ω cable is joined to a piece of 100Ω cable, and the 100Ω cable is terminated by an impedance of 50Ω What is the reflection coefficient for an incident signal travelling along the 50Ω cable at the frequency for which the length of the 100Ω cable is a quarter of a wavelength?

9.6 A mismatched transmission line has a voltage standing wave ratio of 3 and the first minimum of the voltage amplitude is a distance of λ/3 from the end at which it is terminated by an impedance Z. If the characteristic impedance of the line is 200Ω, determine the reflection coefficient at the termination and the complex impedance Z.

9.7 Figure 9.13 shows a schematic plan view of the conductors in a strip line arrangement known as a ‘hybrid ring’. The ring joins together the four strip lines represented by the arms 1, 2, 3 and 4. Each arm has the same characteristic impedance Z0, and the characteristic impedance of the strip line forming the actual ring is . The hybrid ring has the property that there is no reflection for signals arriving along any of the arms if each arm is matched at the end away from the ring. By considering the relative phases of signals travelling around the ring, deduce the outputs along the arms 2, 3 and 4 for an input on arm 1

Figure 9.13.


1
2
3
4
λ/4
λ/4
λ/4
3λ/4

* In practical strip lines edge effects are not negligible, but the simplified expressions given here show roughly the way the line impedances vary with dimensions.

CHAPTER 10

Maxwell’s equations

Four basic laws of electromagnetism have been discussed in the earlier chapters of this volume. These were Gauss’ law; a law expressing the absence of magnetic poles; Ampère’s law relating the magnetic field to the steady currents producing it; and Faraday’s law of induction. We shall show that it is necessary to modify Ampère’s law when the currents change with time and that, when this is done, the four laws make up a set of equations which any electromagnetic field must satisfy everywhere. These equations are called Maxwell’s equations. Much of electromagnetism hereafter in this book consists of finding solutions to Maxwell’s equations which fit the boundary conditions and distributions of charge and current appropriate to different problems. Often one’s physical insight enables approximate solutions to be found which illustrate the most important features of a particular situation. This is the concern of Chapters 11, 12 and 13. In this chapter we will modify Ampère’s law, summarize the earlier parts of the book whilst discussing Maxwell’s equations, and show how the equations are able to predict and describe electromagnetic radiation.

10.1 THE EQUATION OF CONTINUITY

There is a large amount of experimental evidence suggesting that electric charge is a conserved quantity, i.e. that the total electric charge of an isolated system is constant. The best evidence comes from nuclear reactions. For example if calcium nuclei, each with charge 20e, are bombarded with high energy α-particles (with charge 2e) many possible nuclear reactions can take place. However the sum of the charges on any set of reaction products is always equalto 22e. There are no known violations of the principle of conservation of charge. The principle is on a par with the principles of conservation of momentum and energy. On a macroscopic level we have already invoked charge conservation in deducing Kirchhoff's current rule in Chapter 8. We will now use the principle of charge conservation to derive an equation relating the total macroscopic charge and current densities at each point. This equation is called the equation of continuity.

Consider the elementary volume dτ, with sides dx, dy and dz, centred on the point with coordinates (x,y,z), shown in Figure 10.1. If the charge density ρ varies with time, the charge within the volume element at a time t is ρ(x,y,z)dτ, and at time t + dt it is $[\rho + (\partial\rho/\partial t)dt]d\tau$, i.e. the change in the amount of charge within the volume element during the time dt is $(\partial\rho/\partial t)dt\, d\tau$. The charge flowing into the volume element during the time interval dt through the faces A and B, of area dy dz and parallel to the y–z plane, is

$$\left\{ j_x\left(x-\frac{1}{2}dx, y, z, t\right) - j_x\left(x+\frac{1}{2}dx, y, z, t\right)\right\} dy\, dz\, dt$$

$$= -\frac{\partial j_x}{\partial x} dx\, dy\, dz\, dt$$

$$= -\frac{\partial j_x}{\partial x} d\tau\, dt,$$

where jx is the x-component of the current density j. The total charge flowing into the volume element over all six sides during the time dt is

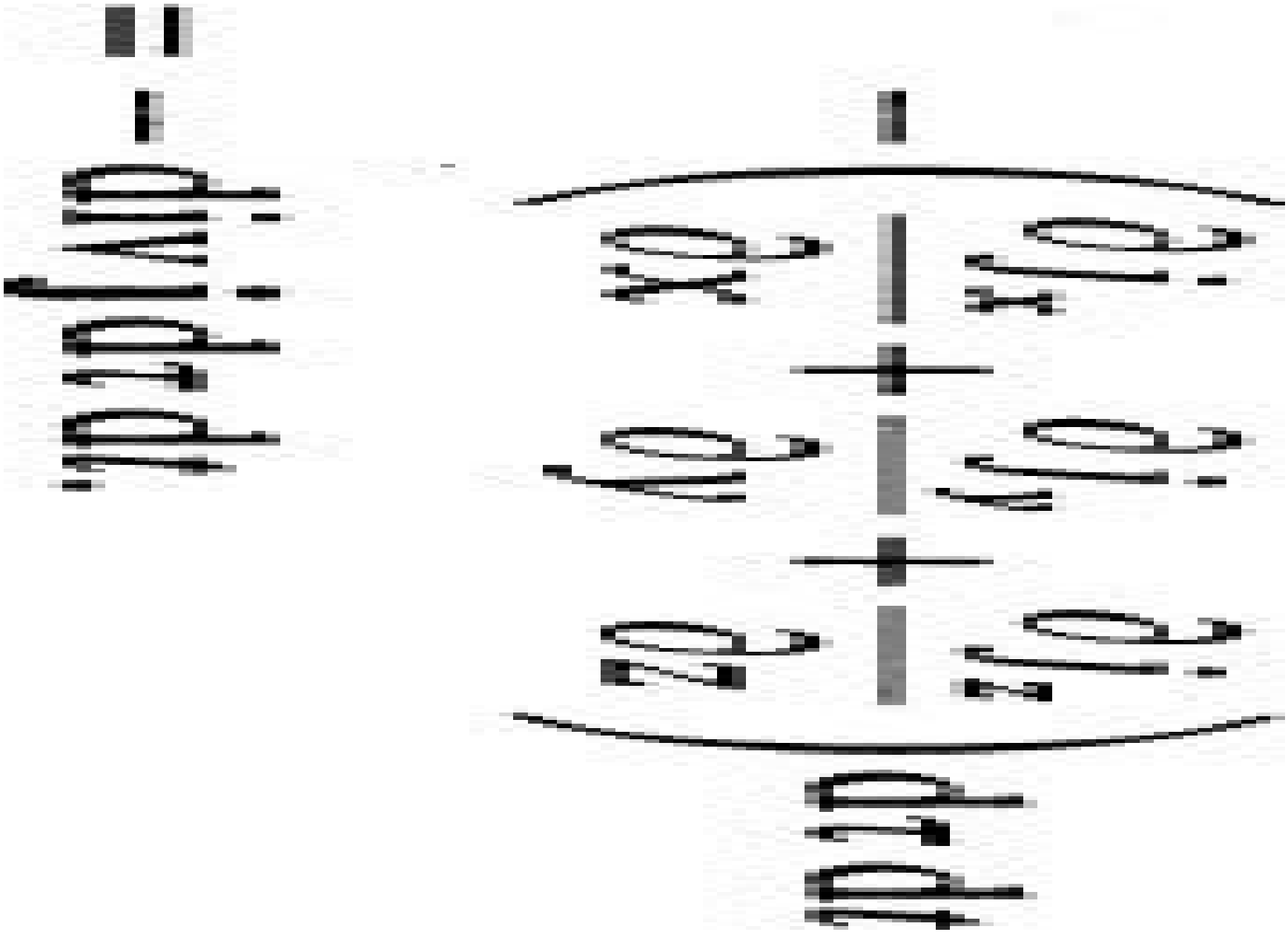
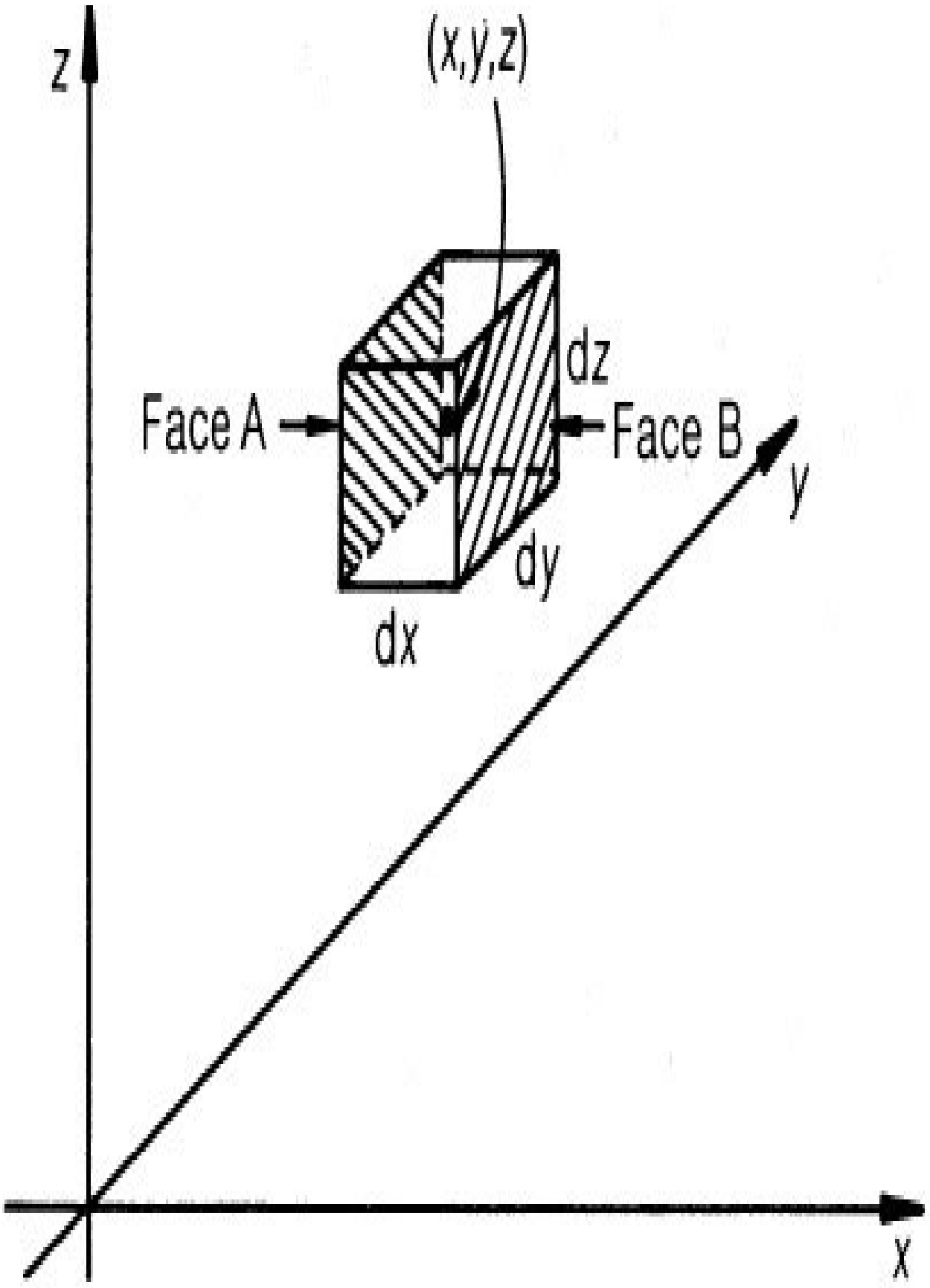
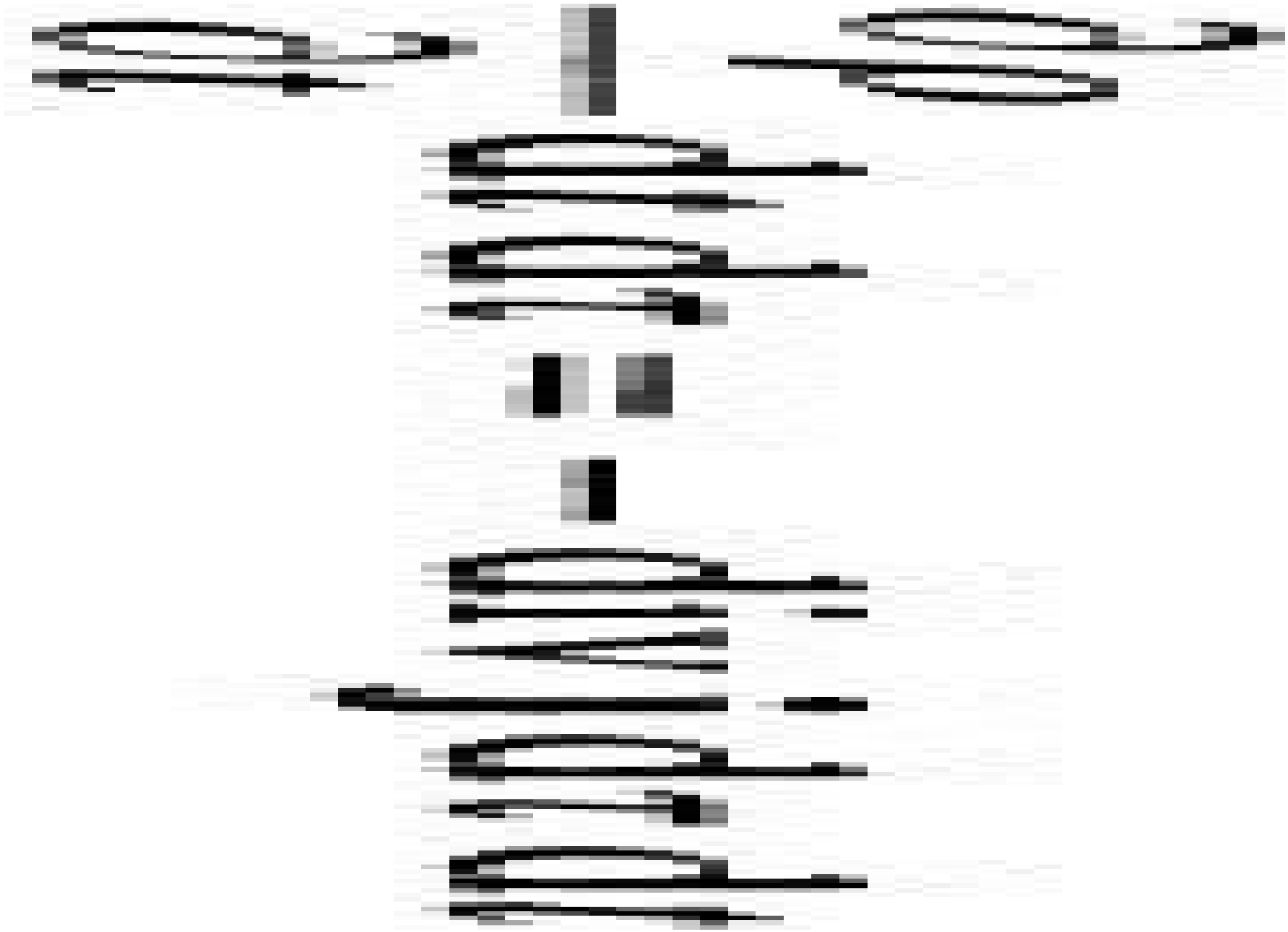


Figure 10.1. An elementary volume element dτ centred on the point (x, y, z).


z
(x,y,z)
dz
Face A
Face B
y
dy
dx
X

Since charge is conserved this charge must equal the change in the total charge within the volume element. Hence



and

(10.1)

The derivation of this equation is independent of the position of the volume element, and so the relation (10.1) holds at all points and at all times. Equation (10.1) is the equation of continuity, expressing the conservation of charge in differential form. We shall now use it to show that the differential form of Ampère's law (as obtained for steady currents) is incompatible with the principle of conservation of charge when the currents are changing with time.

10.2 DISPLACEMENT CURRENT

The differential form of Ampère's law, Equation (4.47), relates the field B at each point to the current density j at the point, and

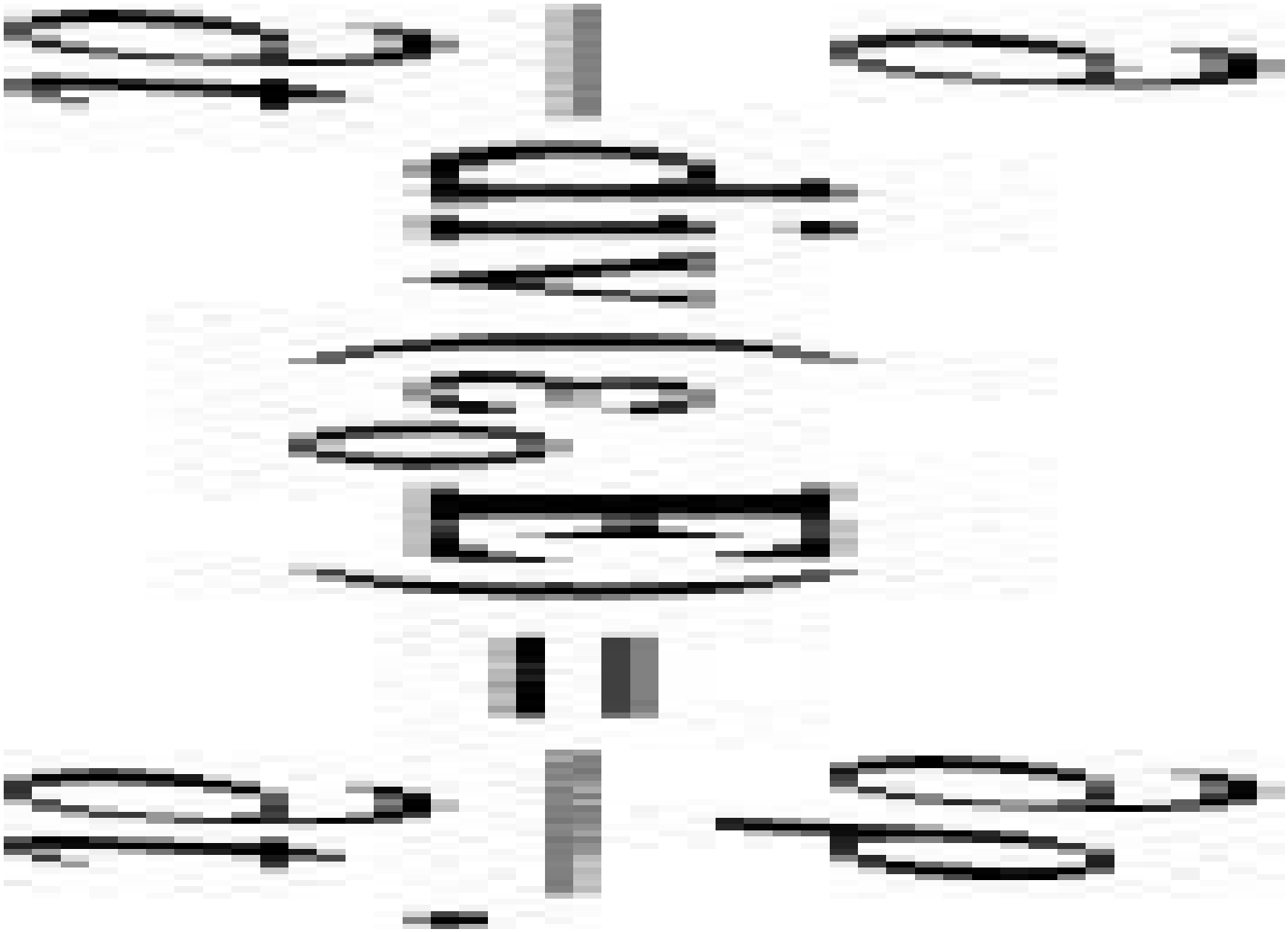
(10.2)

For currents changing with time the divergence of the right-hand side of this equation may be non-zero at some points. For example, if a capacitor is being discharged the charge on the plates is decreasing, and over the surface of the plates $\partial\rho/\partial t$ is non-zero. Hence from Equation (10.1) div j is non-zero. As another example, if an electron beam is scanning across the front face of a television tube, the volume density of charge within the tube is changing with time. If the divergence of the right-hand side of Equation (10.2) can be non-zero, the equation is clearly inadequate for changing currents since the divergence of the left-hand side is identically zero everywhere. The function obtained by taking the curl of another vector function can easily be shown to have zero divergence at all points. We thus see that Ampère's law, as we have used it so far, is not true for currents varying with time. The law has to be modified by adding to the right-hand side of Equation (10.2) a quantity which will make the divergence of the new right-hand side everywhere zero. This requirement is satisfied, according to the equation of continuity, if the divergence of the quantity we add is equal to $\partial\rho/\partial t$.

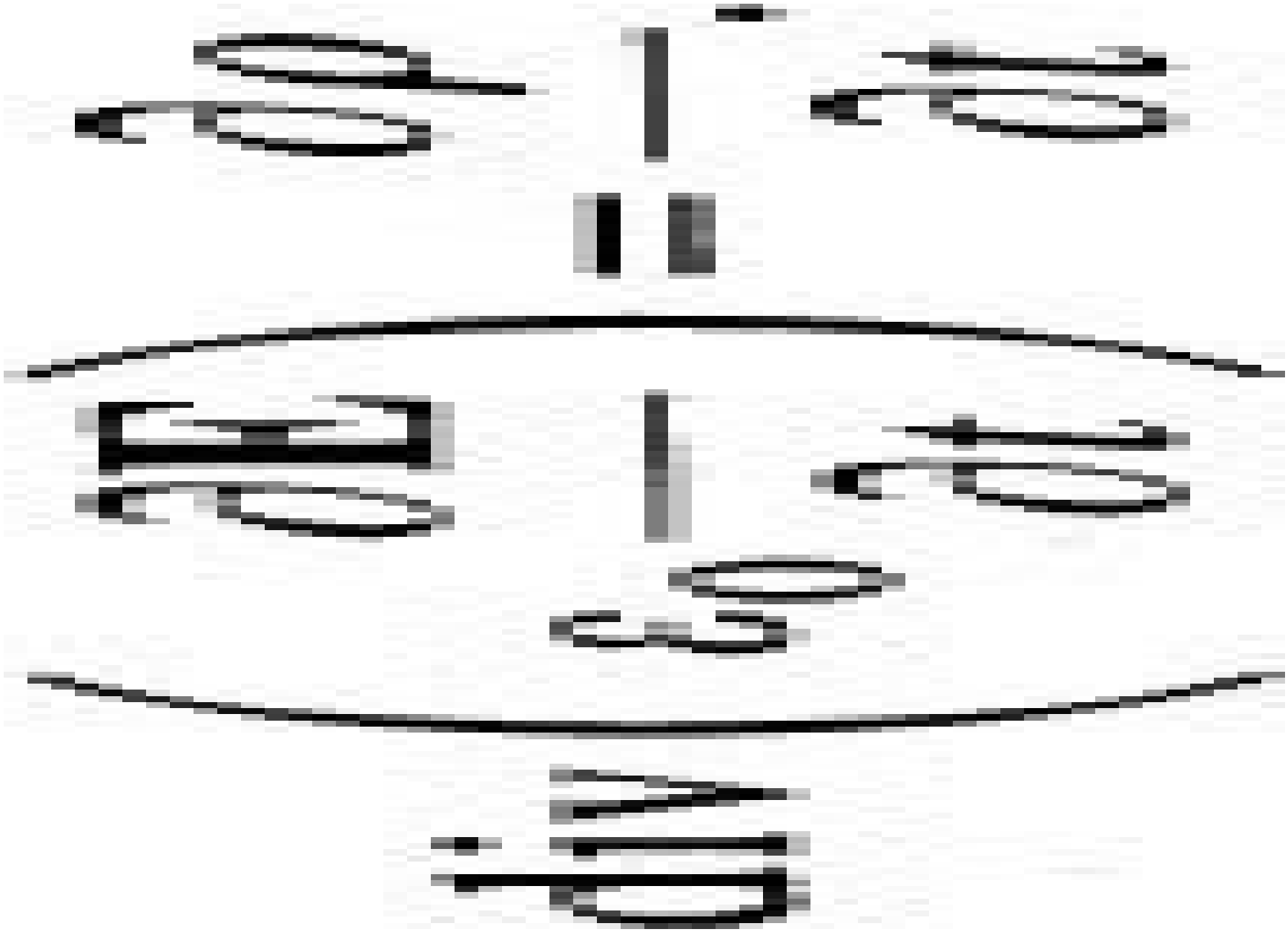
Let us examine the differential form of Gauss' law obtained in section 1.4.4,

(10.3)

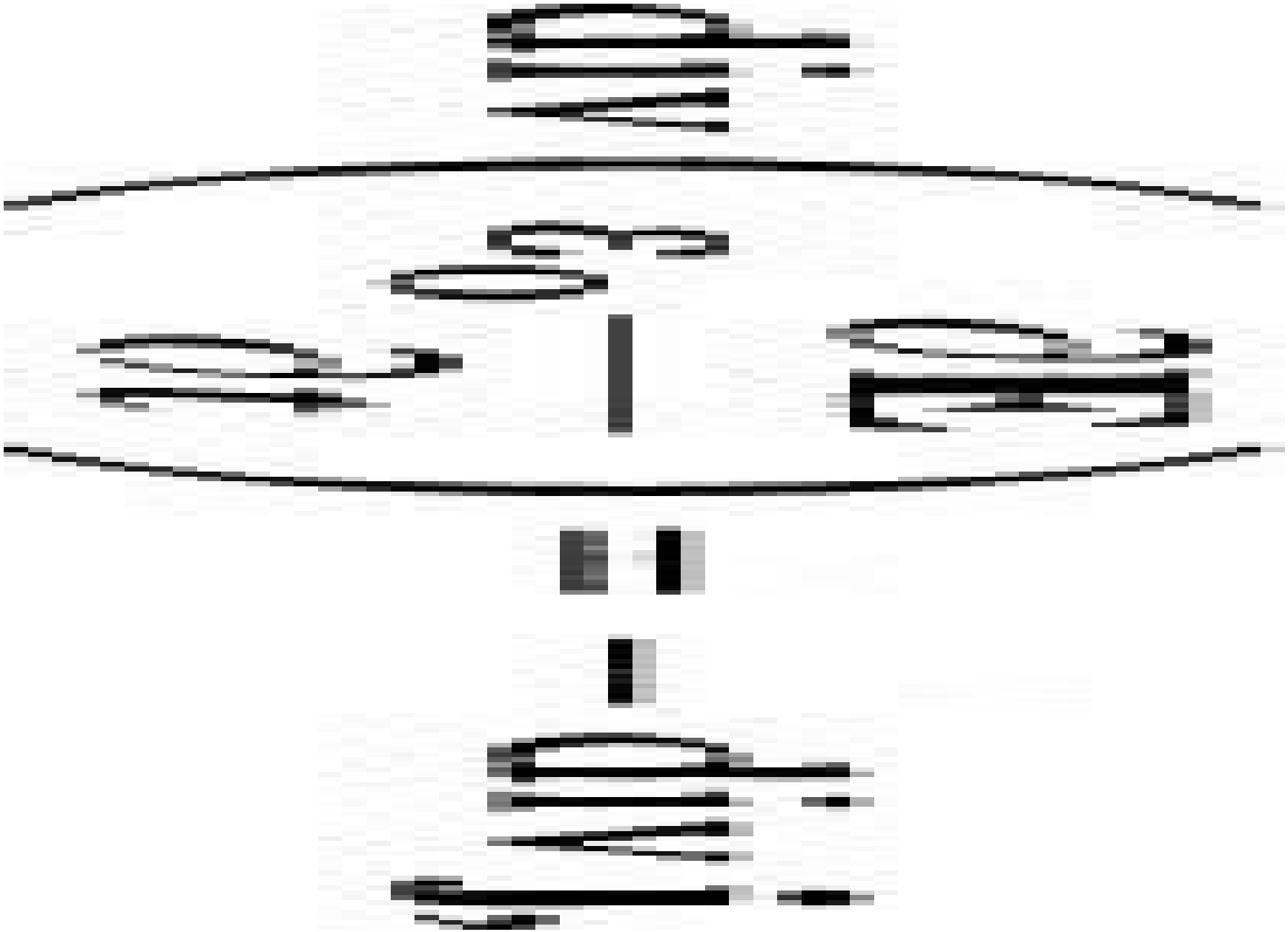
If the charge density ρ changes with time the changing magnetic fields produced by the varying currents will induce an electric field. However the divergence of the induced electric field is everywhere zero, and so Equation (10.3) remains valid for fields that change with time. Taking the derivative with respect to time of this equation,



The order of differentiation on the left-hand side can be changed, since the operator div does not depend on time, giving



Hence

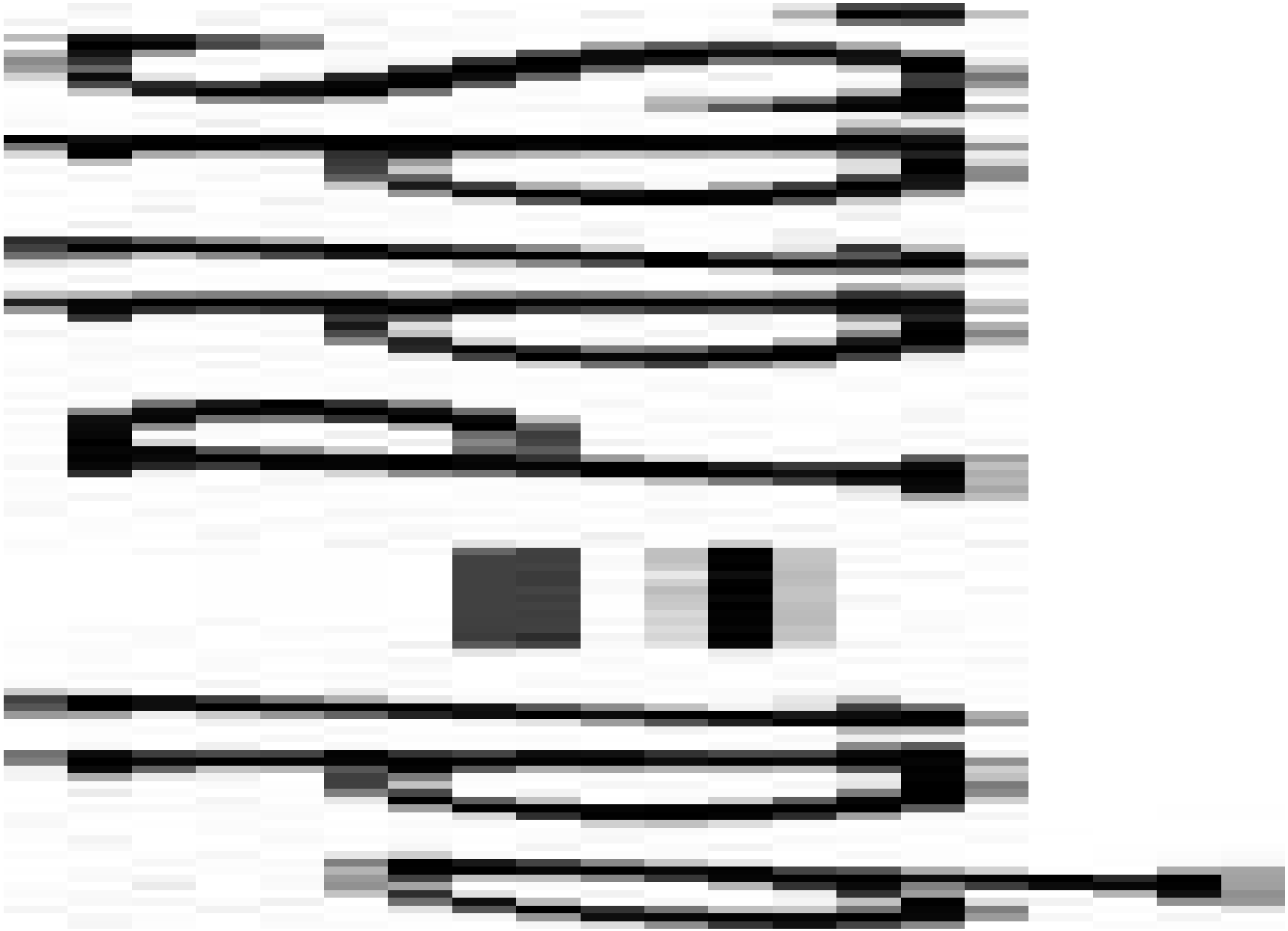


from the equation of continuity. Hence adding the quantity $\varepsilon_0 \partial E/\partial t$ to the right-hand side of Equation (10.2) gives the required modification to Ampère's law, which becomes

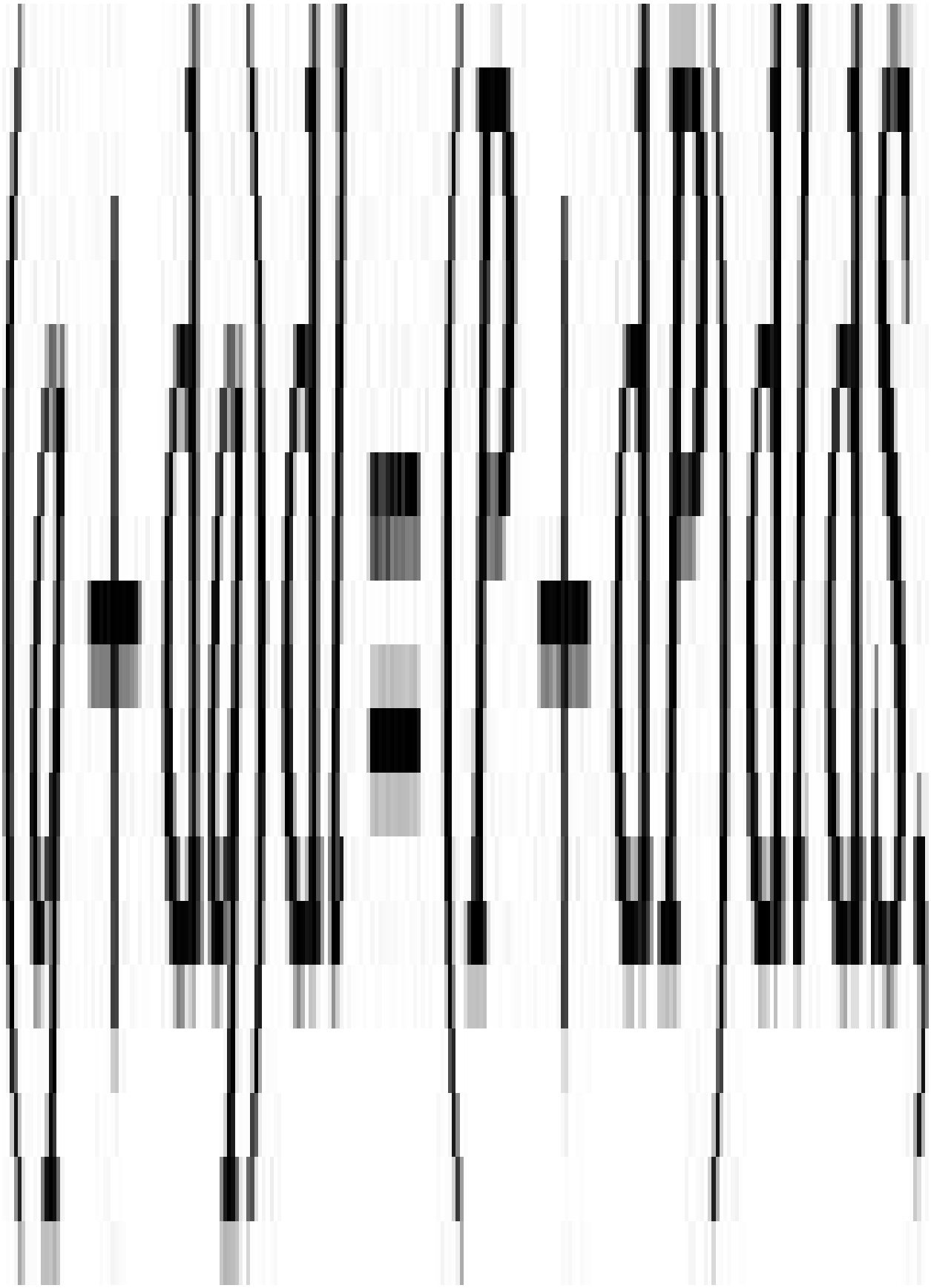
(10.4)

This equation is always true, but only has a simple form in the absence of materials. In the presence of matter the total current density j may have contributions from the medium itself. In that situation it is difficult to use the equation as it stands; it is easier to use equations expressing relationships between fields and free charges and current densities, which can be specified in a simple way. One possible contribution to the current density, already discussed in section 5.2, arises from the magnetization. A non-uniform magnetization M is equivalent to a current density j_M given by Equation (5.19),

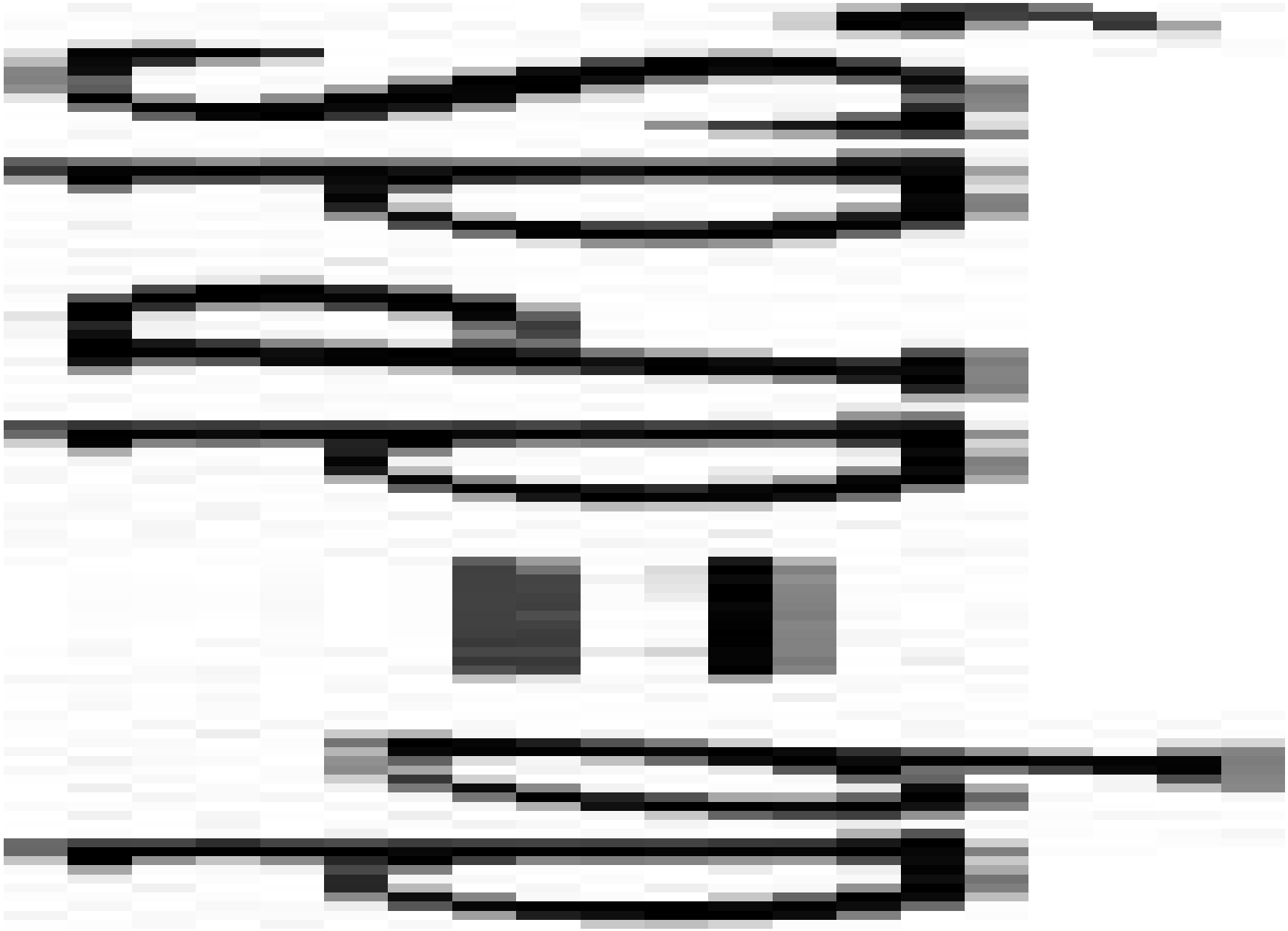
Now that we are considering fields which change with time a second contribution to the total current density from matter may arise. This contribution is present when the electric polarization P of a medium changes with time, the resulting changes in dipole moments corresponding to motions of charge, i.e. currents, through any area within the dielectric. These currents can be associated with a polarization current density, jP, at every point within the medium. To determine the relationship between jP and the polarization consider, at any point in the medium, a small volume element whose shape is a right cylinder of height dl with end caps of area dS. Let the axis of the cylinder be in the direction of thepolarization at the point. At time t when the polarization is P the electric dipole moment of the small volume has size P dl dS and we can imagine this to be due to charges ±q on the end caps, so that



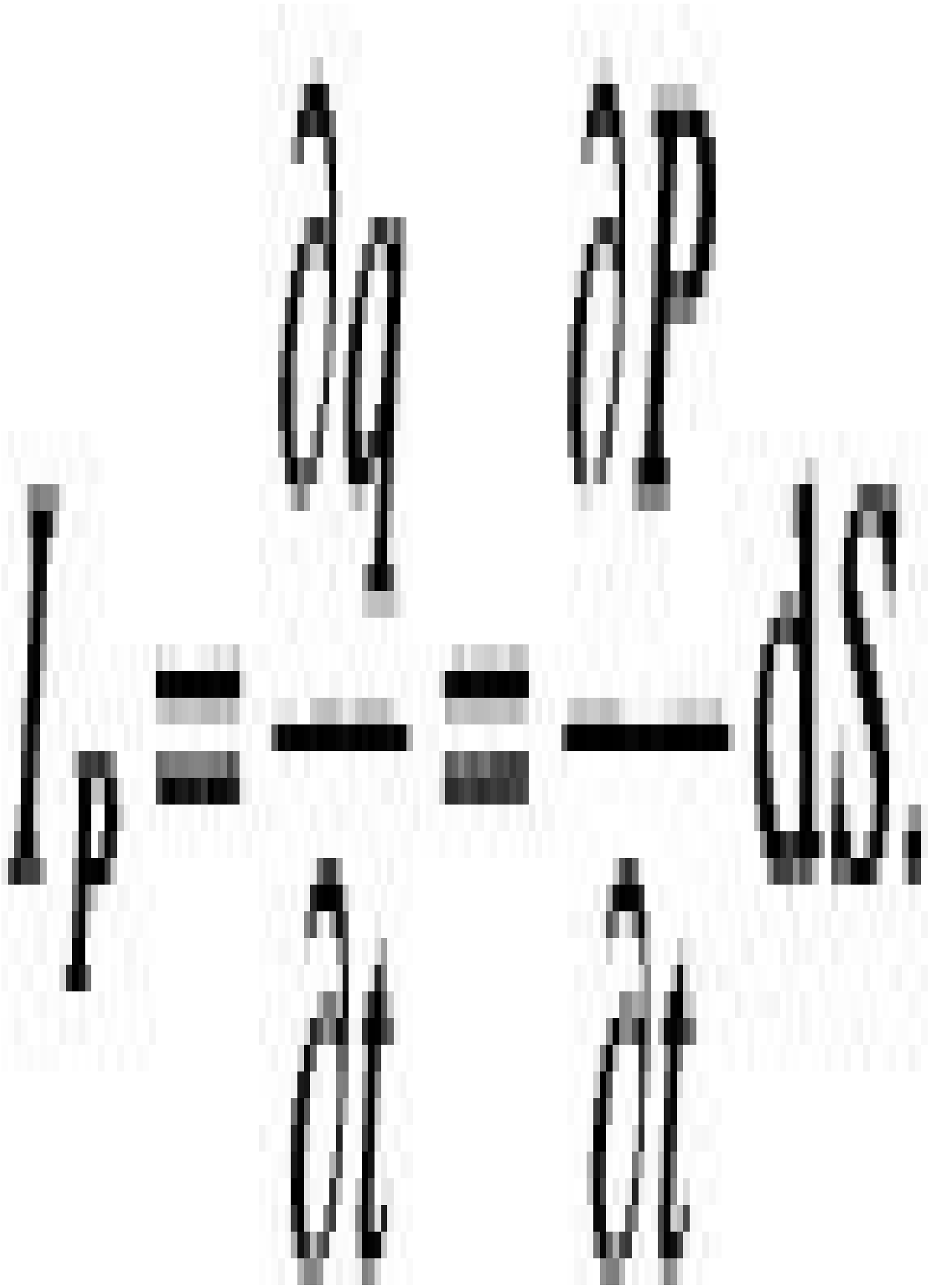
At time t + dt the polarization has changed to (P + dP) and



so that



and the current IP across an area dS around the point is equal to



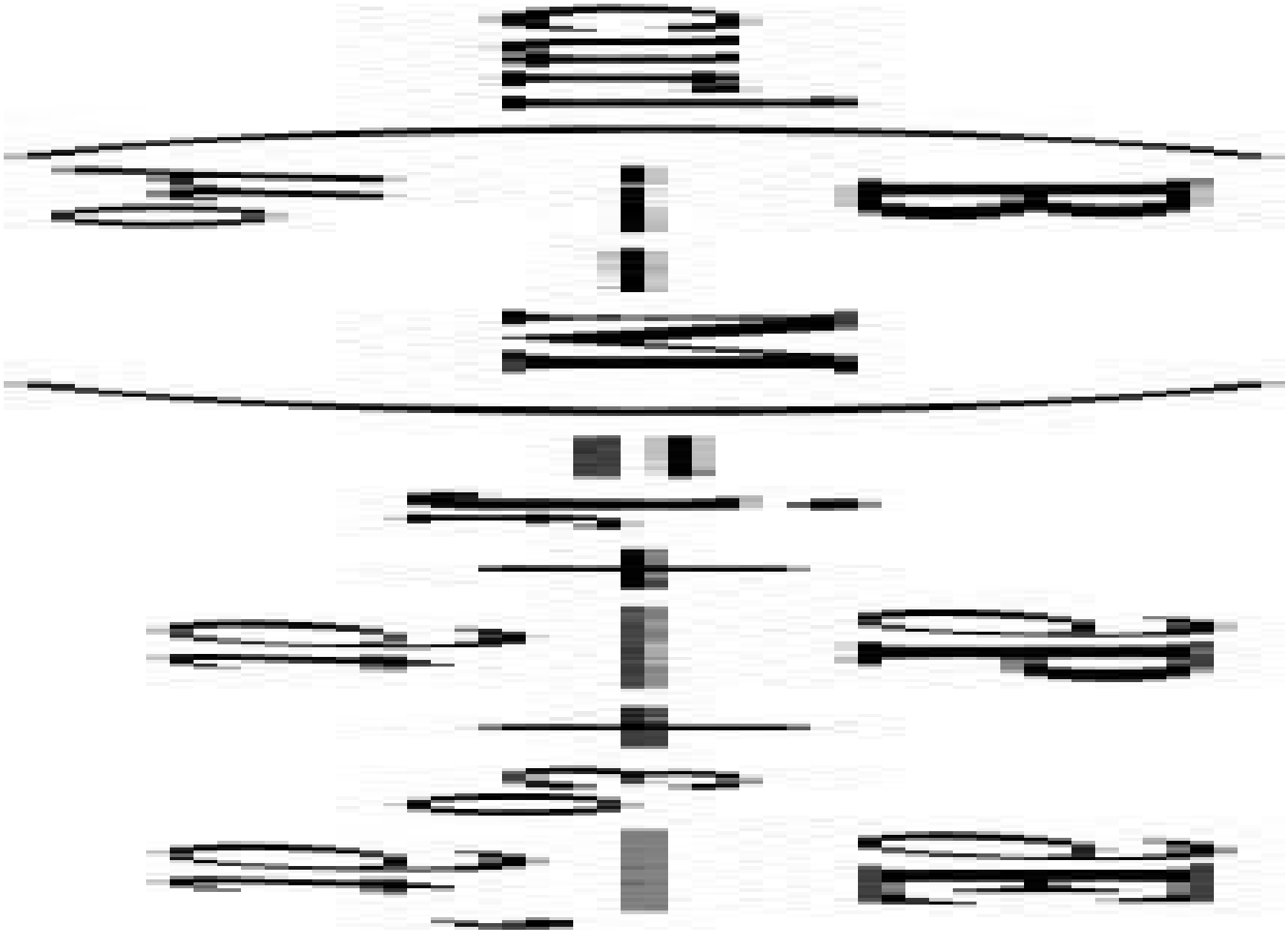
The current flow is in the direction of the vector dP and hence the polarization current density is given by

(10.5)

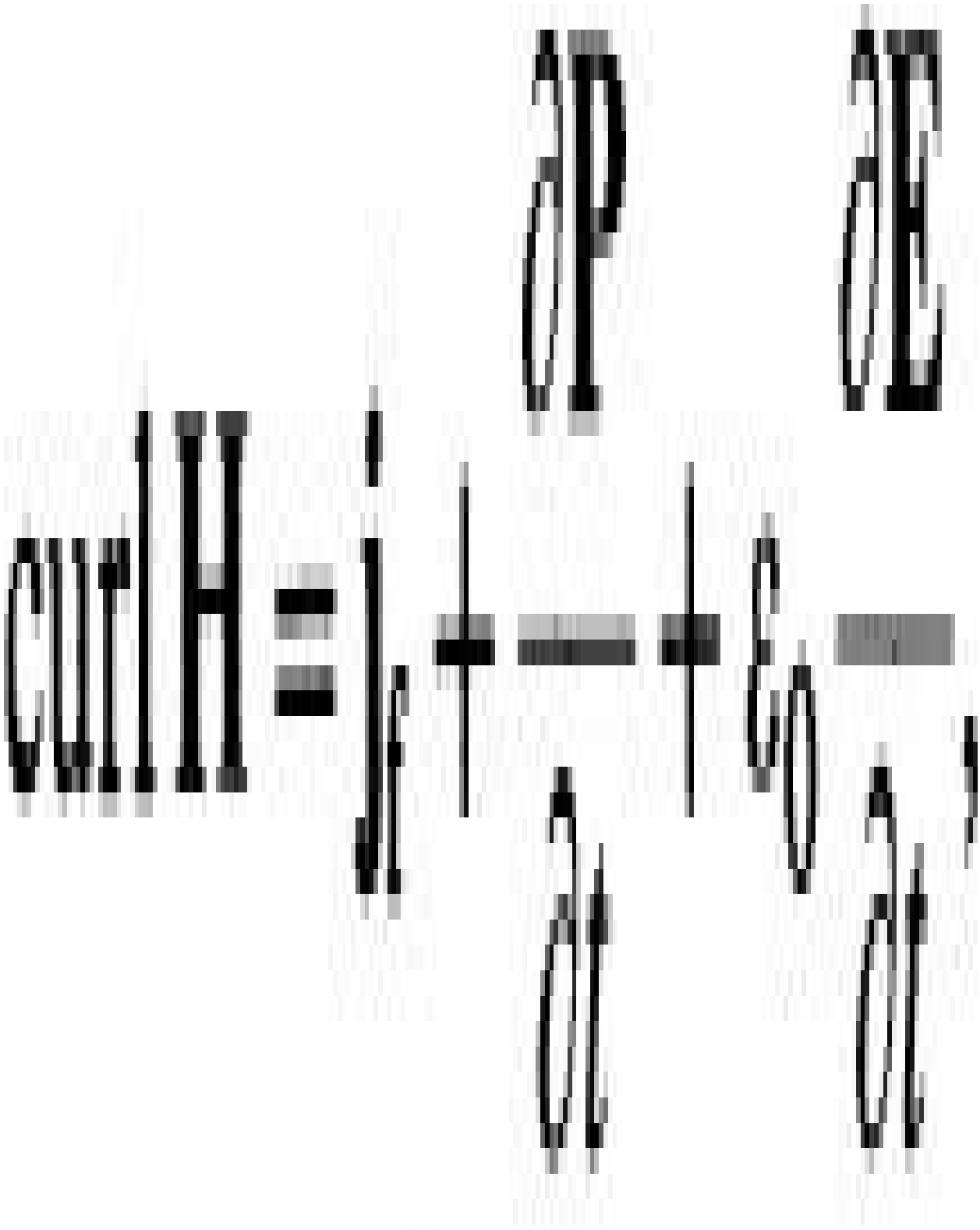
Rewriting the total current density j in Equation (10.4) in terms of the separate contributions jf, jM and jP we have,

(10.6)

Now writing the magnetization current density in terms of the magnetization and the polarization current density in terms of the polarization, and rearranging, Equation (10.6) becomes



or



where H is the magnetic intensity defined by Equation (5.27). The last two terms on the right-hand side of this equation are equal to the time derivative $\partial D/\partial t$ of the field D defined by Equation (2.20). With this change we finally obtain, in terms of the fields H and D and the free current density j_f, the generalization of Ampère's law which satisfies the equation of continuity. It is

(10.7)

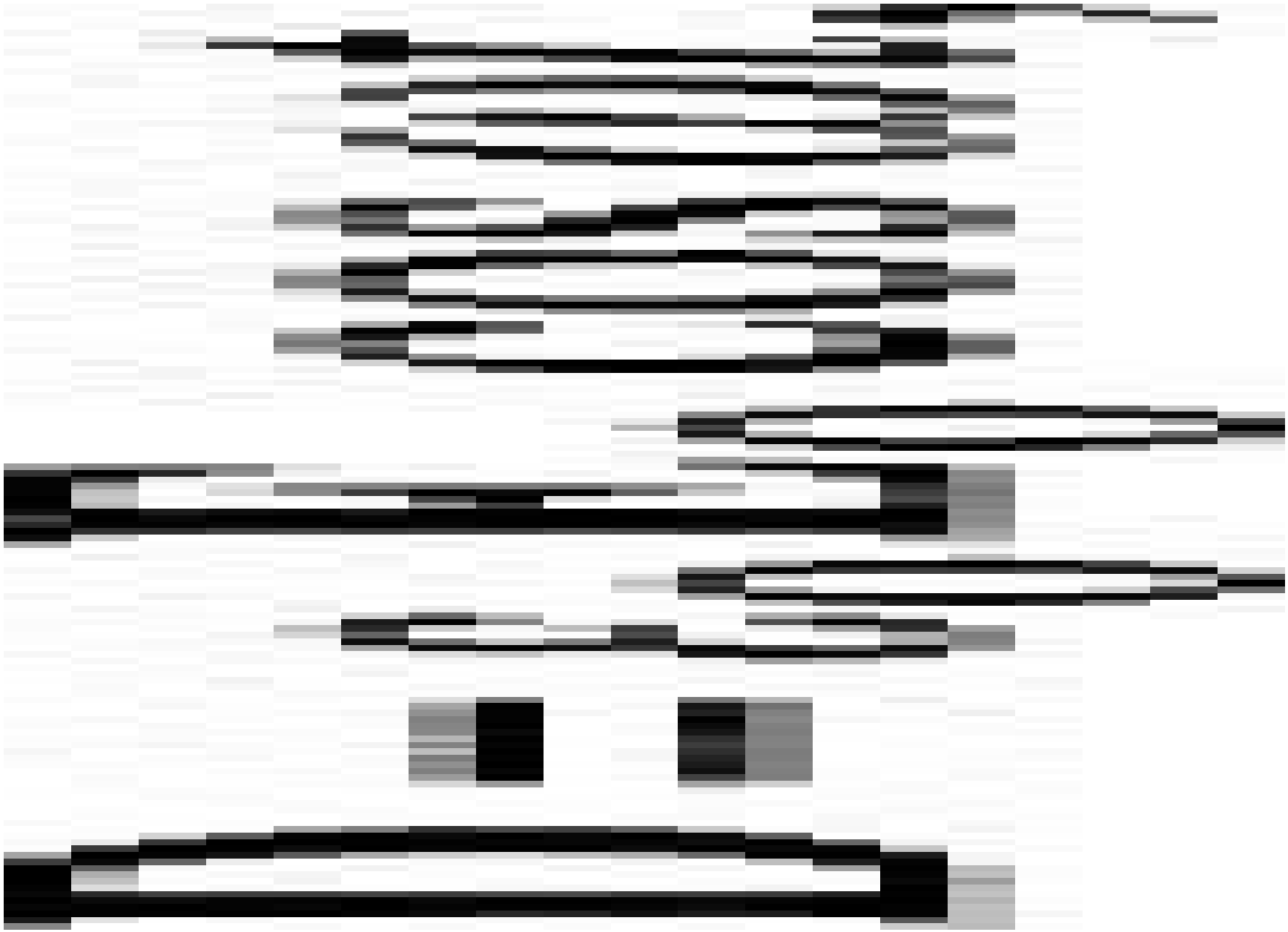
The term $\partial D/\partial t$ is called the displacement current density. It may have two different components. One occurs when a polarization changes with time, and can easily be visualized, for example, as a current flowing back and forth between the plates of a capacitor filled with a polarizable medium as the charge distributions within the atoms vary with the field across the plates. The other component is always present, even if media are absent.

Displacement current was introduced by Maxwell in 1862. With the benefit of hindsight it seems surprising perhaps that people should have thought Ampère's original law would apply to non-steady currents. However the original law correctly described the results of all the experiments made with the slowly changing currents which were met in the laboratory at the time of Maxwell. 'The form of the rules we write down is more general than the experience from which they are culled, and it is quite normal to apply them to situations we have not directly experienced. The act of genius is to realize precisely where it is that we are going too far*.' Where nowadays are we making the same sort of error as that to which Maxwell drew attention in electromagnetism?

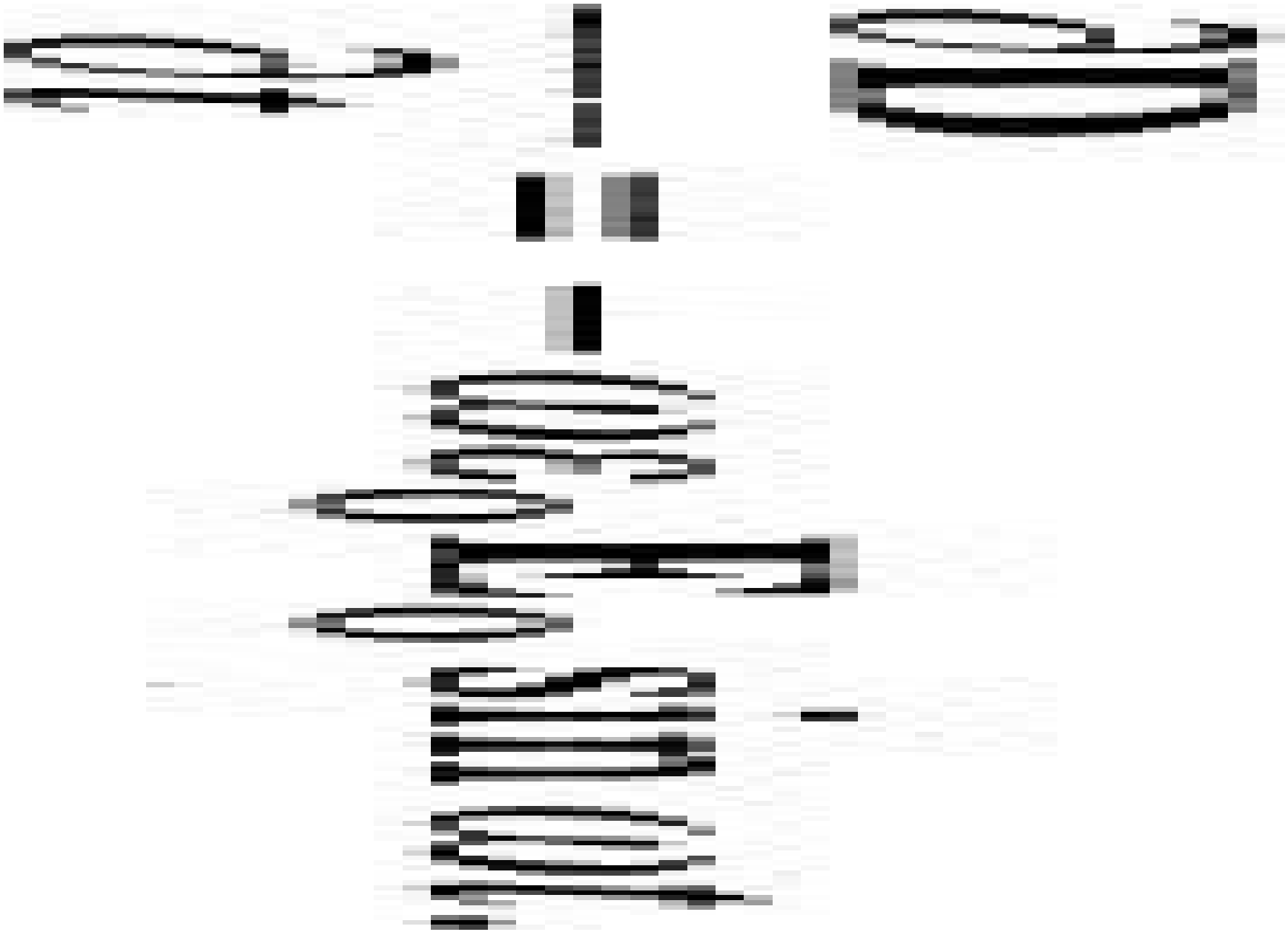
The inclusion of displacement current in the electromagnetic field equations restores a degree of symmetry to electricity and magnetism. A changing magnetic field produces an electric field according to Faraday's law. Now we see that a changing electric field produces a magnetic field. This fact has far-

reaching consequences.

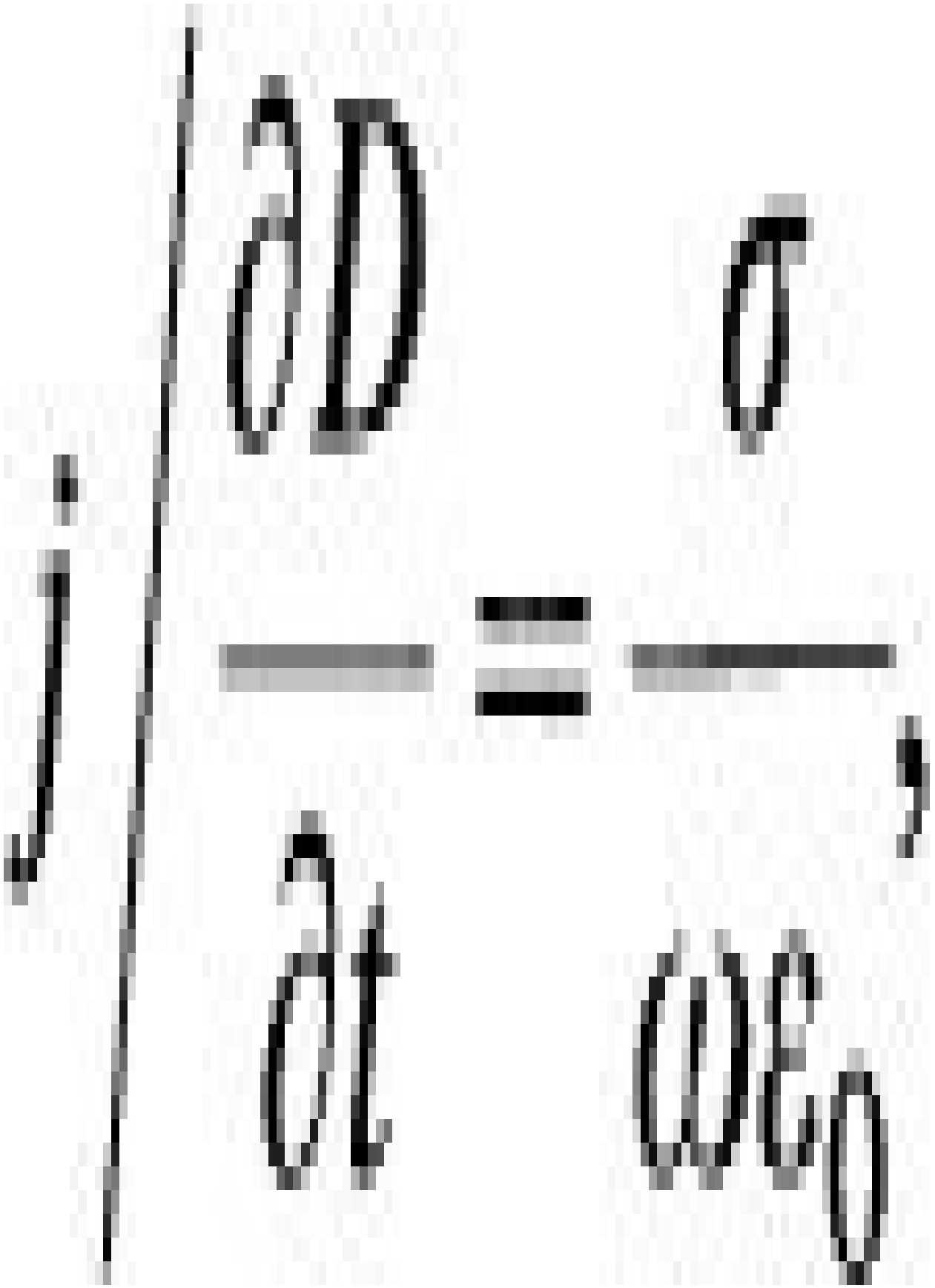
Displacement currents are not usually important compared to the currents arising from the motion of free charges when investigating the behaviour of the continuous circuits met in the laboratory and discussed in Chapters 7 and 8. In order to obtain an idea of the relative size of the two types of current in conductors, let us calculate them for a wire in which there is an alternating electric field $E_0 \cos \omega t$, where ω is the angular frequency. The current density j is given by $j = \sigma E_0 \cos \omega t$ where σ is the electrical conductivity, which is to a high degree of accuracy equal to the steady current conductivity up to the highest frequencies used. If we assume that there is no contribution to the macroscopic electric field from the atoms of the metal, the relative permittivity of the metal is unity and the electric displacement D inside the wire is given by



and

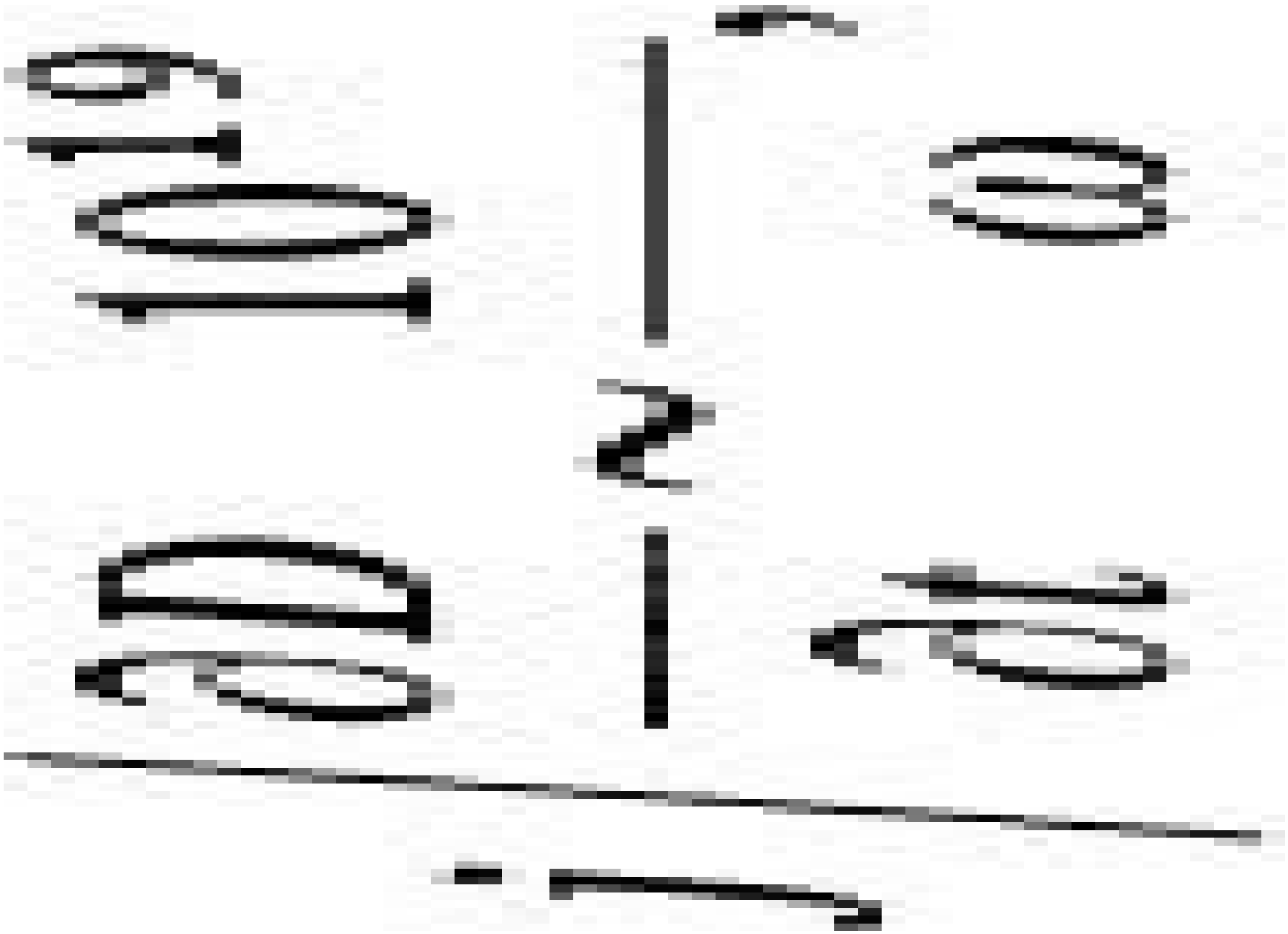


Therefore



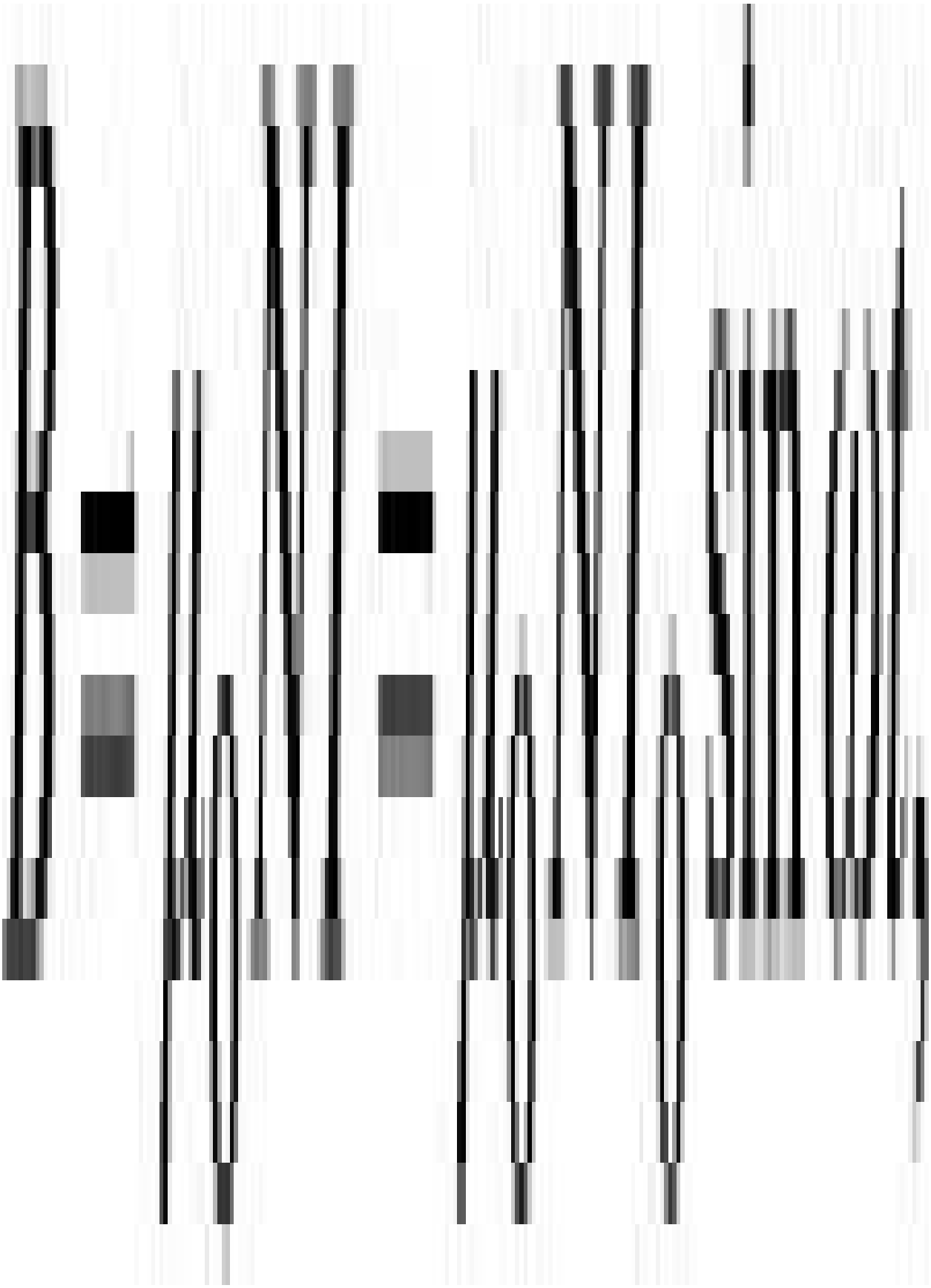
where j and $\partial D/\partial t$ are the amplitudes of the corresponding vectors.

For copper we obtain



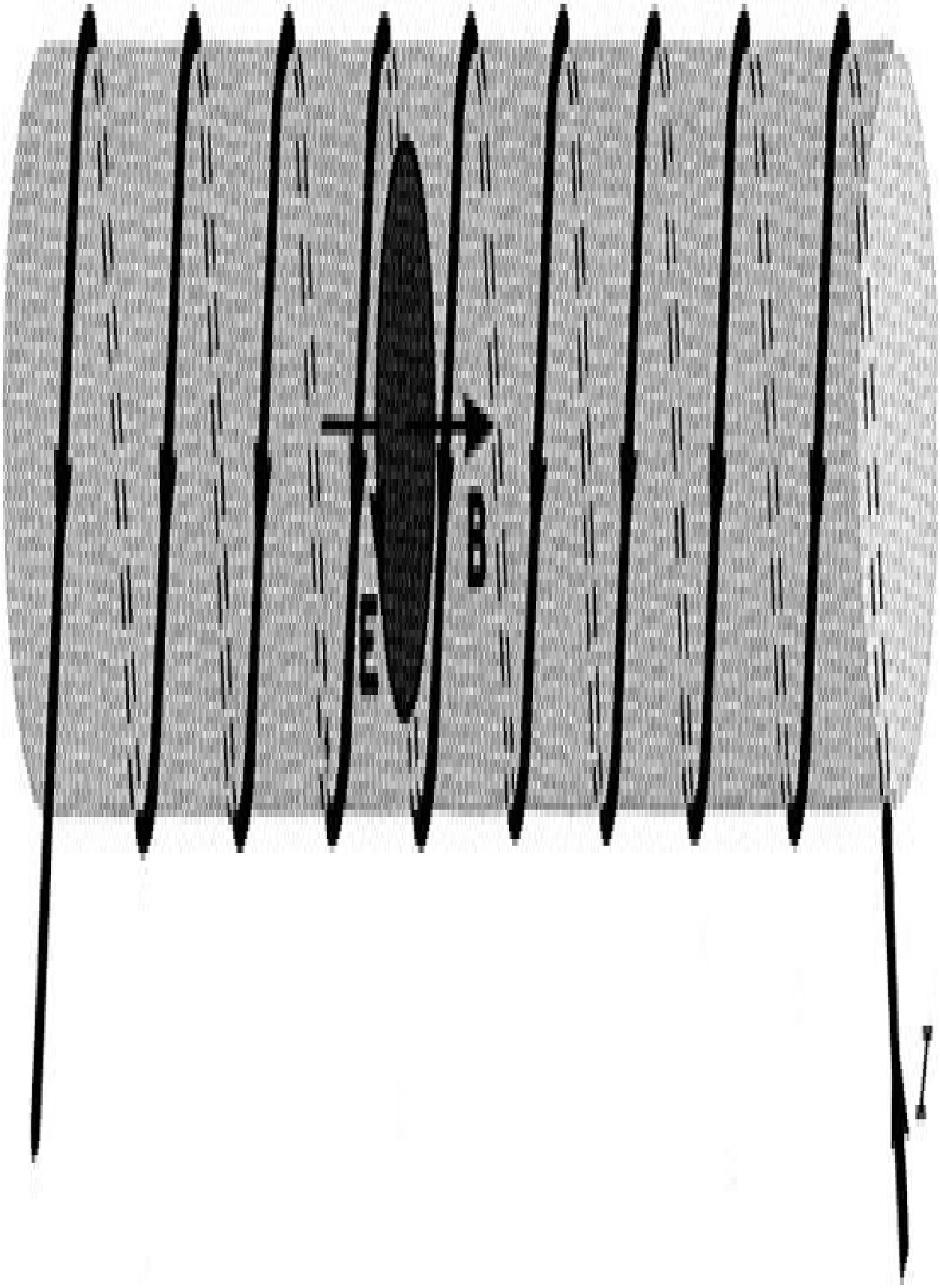
and the ratio of the conduction current to the displacement current in a wire is very large at all frequencies used in electrical circuits.

Since the displacement current density is non-zero outside the conductor the above calculation is not directly relevant to the question of the relative importance of conduction and displacement currents for the magnetic fields produced by changing currents. In order to investigate this point we will discuss the particular example of a long solenoid carrying an alternating current $I = I_0 \sin \omega t$, where ω is the angular frequency at which the current, of amplitude I_0, oscillates. The solenoid has N turns per unit length and radius a. Let us first neglect the displacement current. Inside the solenoid the magnetic field B can be determined as in section 4.4.3, and is to a good approximation parallel to the axis and everywhere constant. Its magnitude is given by

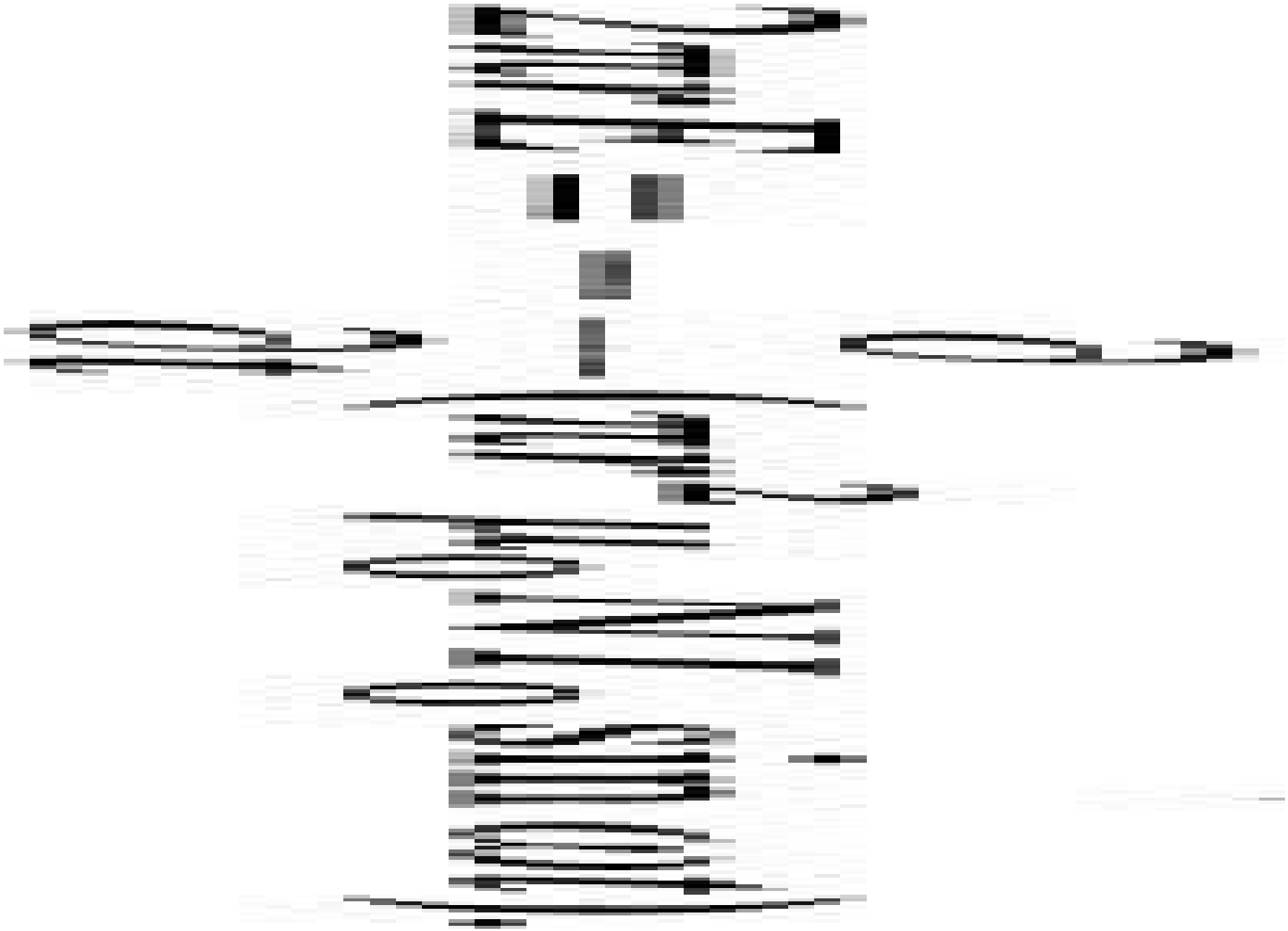


from Equation (4.36). The magnetic field changes with time and hence there is an induced electric field. Inside the solenoid, by symmetry, all points on any circle with centre on the axis and plane perpendicular to the axis are indistinguishable. The induced electric field must therefore have constant magnitude at all points on the circle and be in a direction tangential to the circle, as shown in Figure 10.2. Its direction is given by Lenz's law. Its magnitude

Figure 10.2. The electric and magnetic fields inside a solenoid carrying an alternating current. If the field B is decreasing in the direction shown, the electric field is anticlockwise looking into the solenoid from the right. The amplitude of the electric field is proportional to the distance from the centre (Equation (10.8)).


B
E

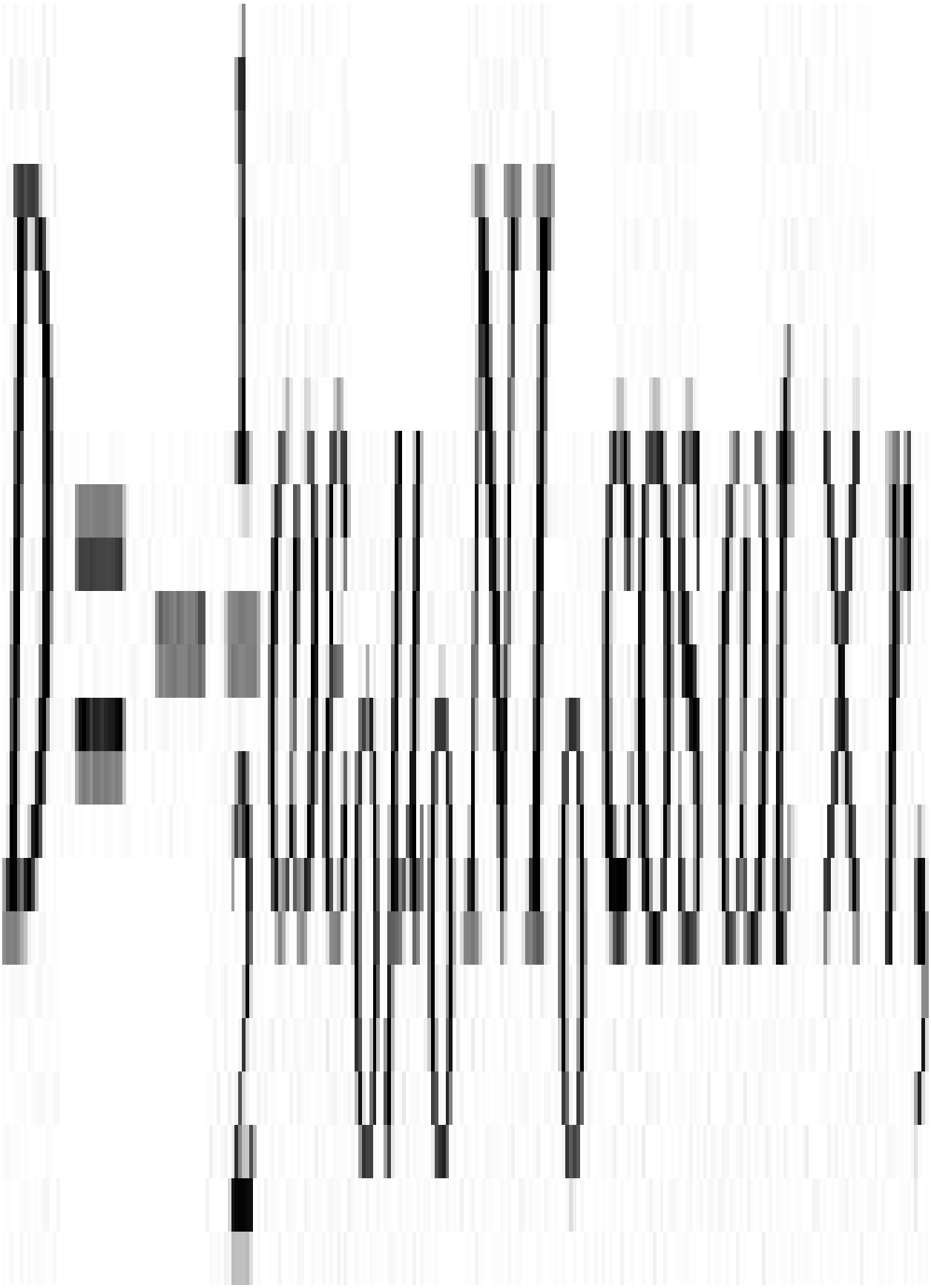
E can be determined by applying Faraday's law to a circle of radius r:



Hence

(10.8)

The electric displacement D is given by

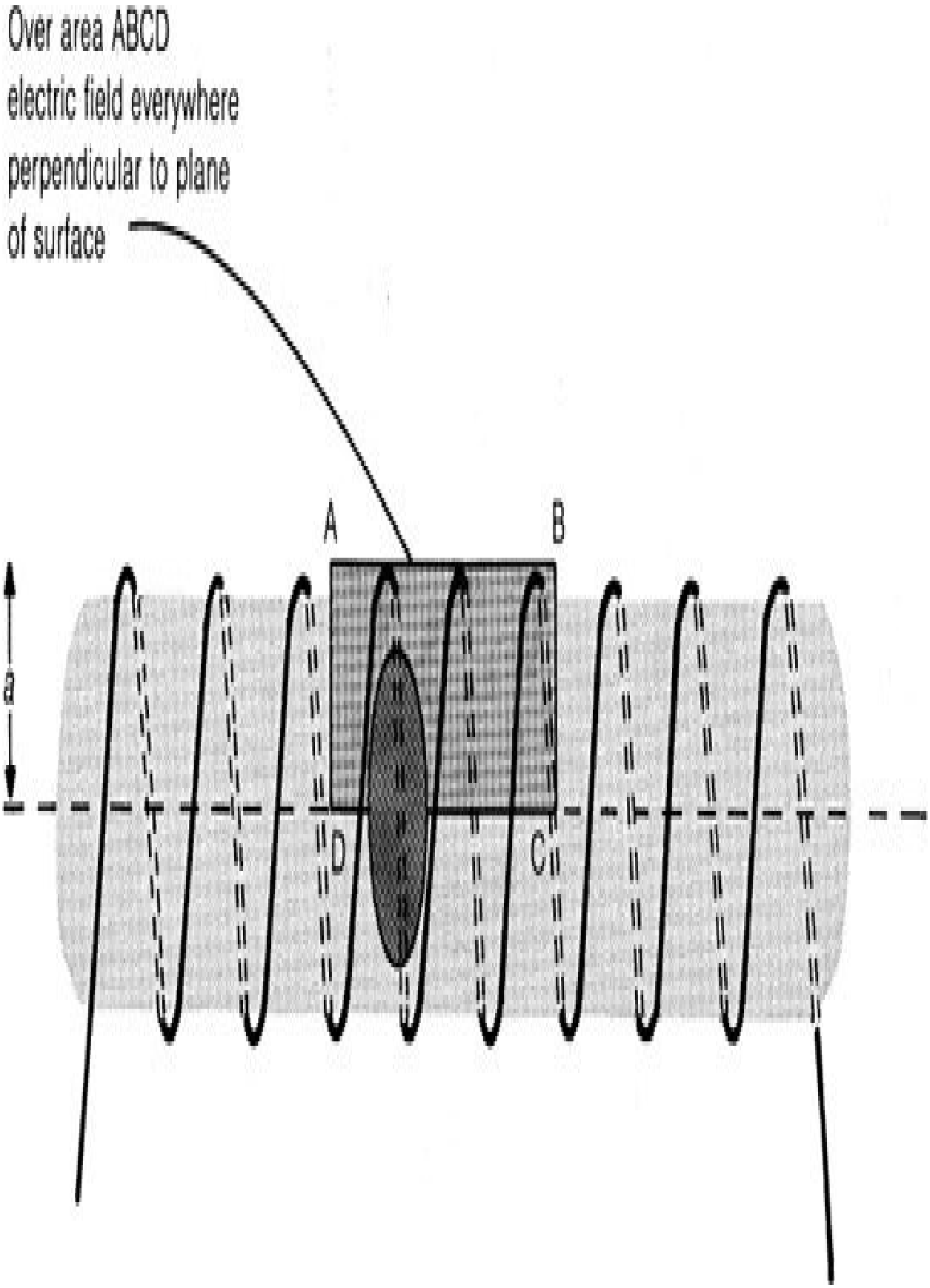


where we have put the relative permittivity of the air in the solenoid equal to unity. Differentiating this equation with respect to time we obtain

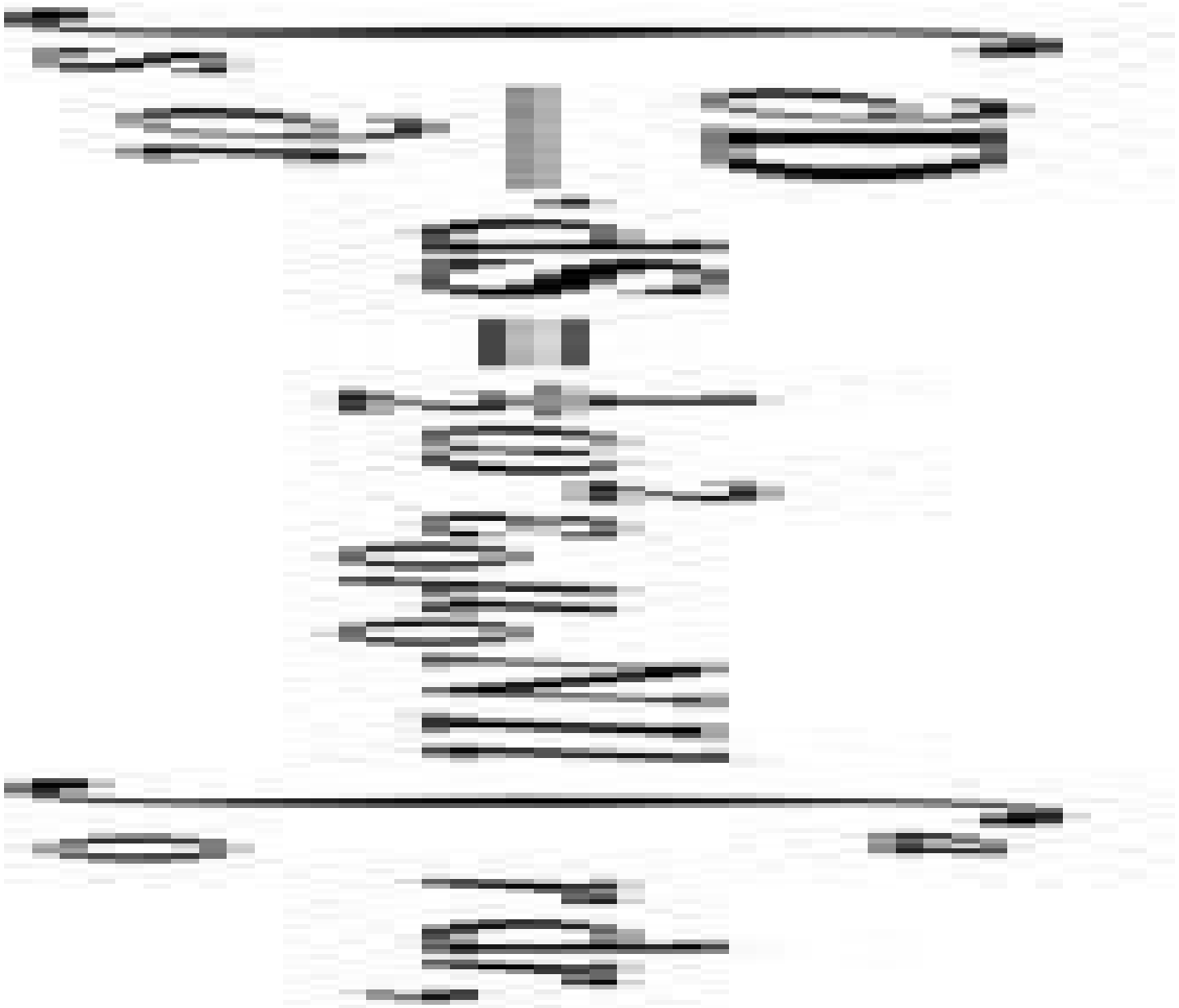
(10.9)

We are now in a position to determine whether the original assumption of negligible displacement current was justified. The magnetic field inside the solenoid can be worked out by applying Ampère's integral relationship to a rectangular loop like the one shown in Figure 10.3. The current through the surface S of the loop i.e. the surface integral $\int s j \cdot dS$, is equal to NIl. The displacement current, which was neglected, is equal to $\int s (\partial D/\partial t) \cdot dS$, and an estimate of the error involved in its neglect is given by the ratio

Figure 10.3. The current due to the motion of free charges and the displacement current through the surface ABCD


Over area ABCD
electric field everywhere
perpendicular to plane
of surface
A
B
a
D
C

From Equation (10.9),



since over the surface of the loop the vector $\partial D/\partial t$ is everywhere parallel to the vector dS. Hence the ratio of the displacement current through the loop to the conduction current is about If the current alternates at a frequency of 1 MHz and the radius is 5 cm, the numerical value of this ratio is ~10^{-6}. This shows that the contribution from the displacement current to the magnetic field inside the solenoid is very small. It does not mean that there are no important effects due to displacement currents at frequencies of the order of 1 MHz, neither does it mean that the electric field inside the solenoid is negligible. If the amplitude of the current is 1A and the coil has 50 turns per metre, the amplitude of the electric field near to the coil, given by Equation (10.8), is about 10 Vm^{-1} at 1 MHz.

10.3 MAXWELL'S EQUATIONS

We can now summarize the equations which the field vectors E, D, B and H everywhere satisfy. These equations are called Maxwell's equations. They are:

(10.10) (10.11) (10.12) (10.13)

Any electromagnetic field must satisfy all of Maxwell's equations.

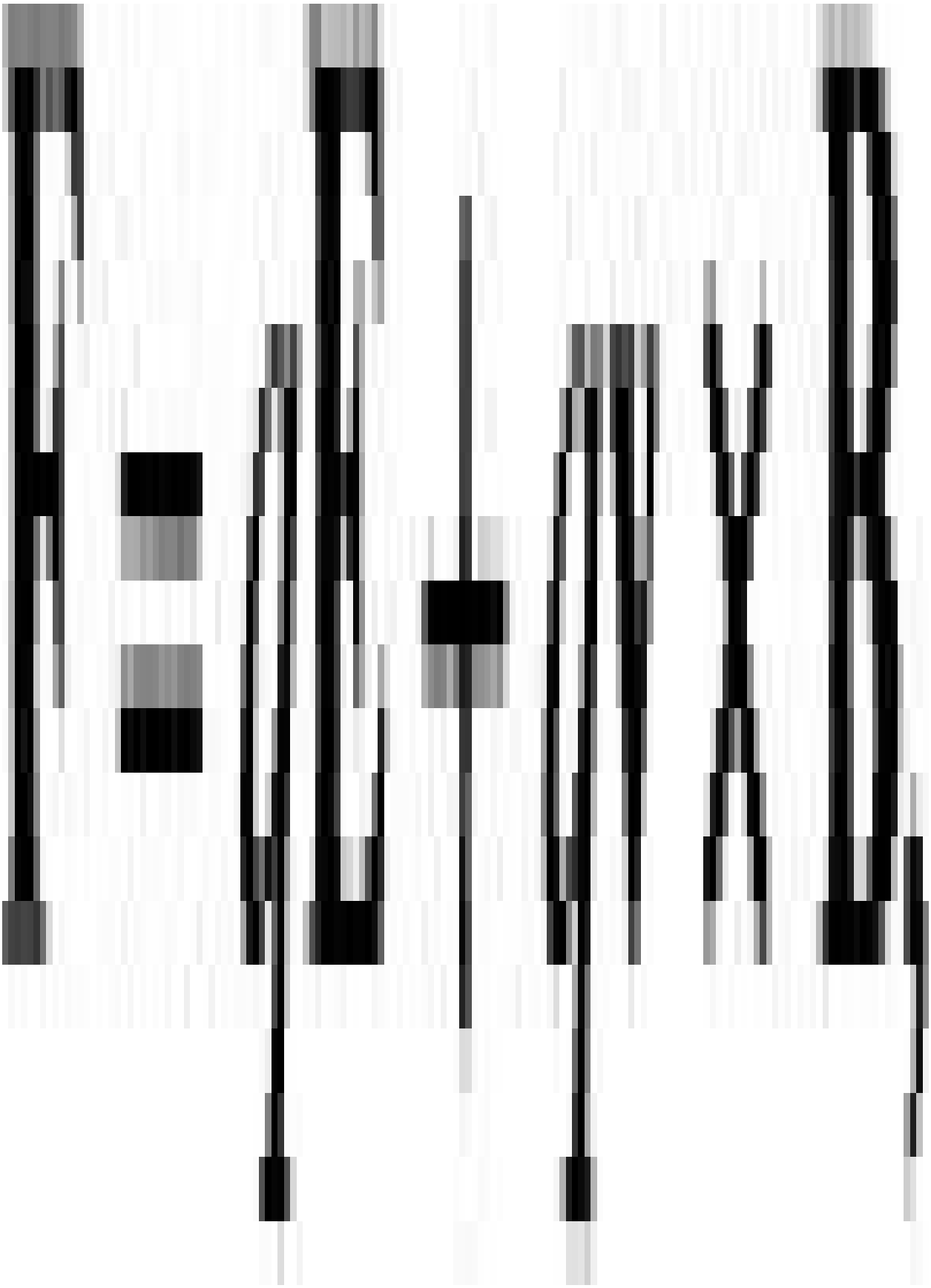
For the isotropic and homogeneous materials considered in this book the equations

$$D = \varepsilon \varepsilon_0 E,$$

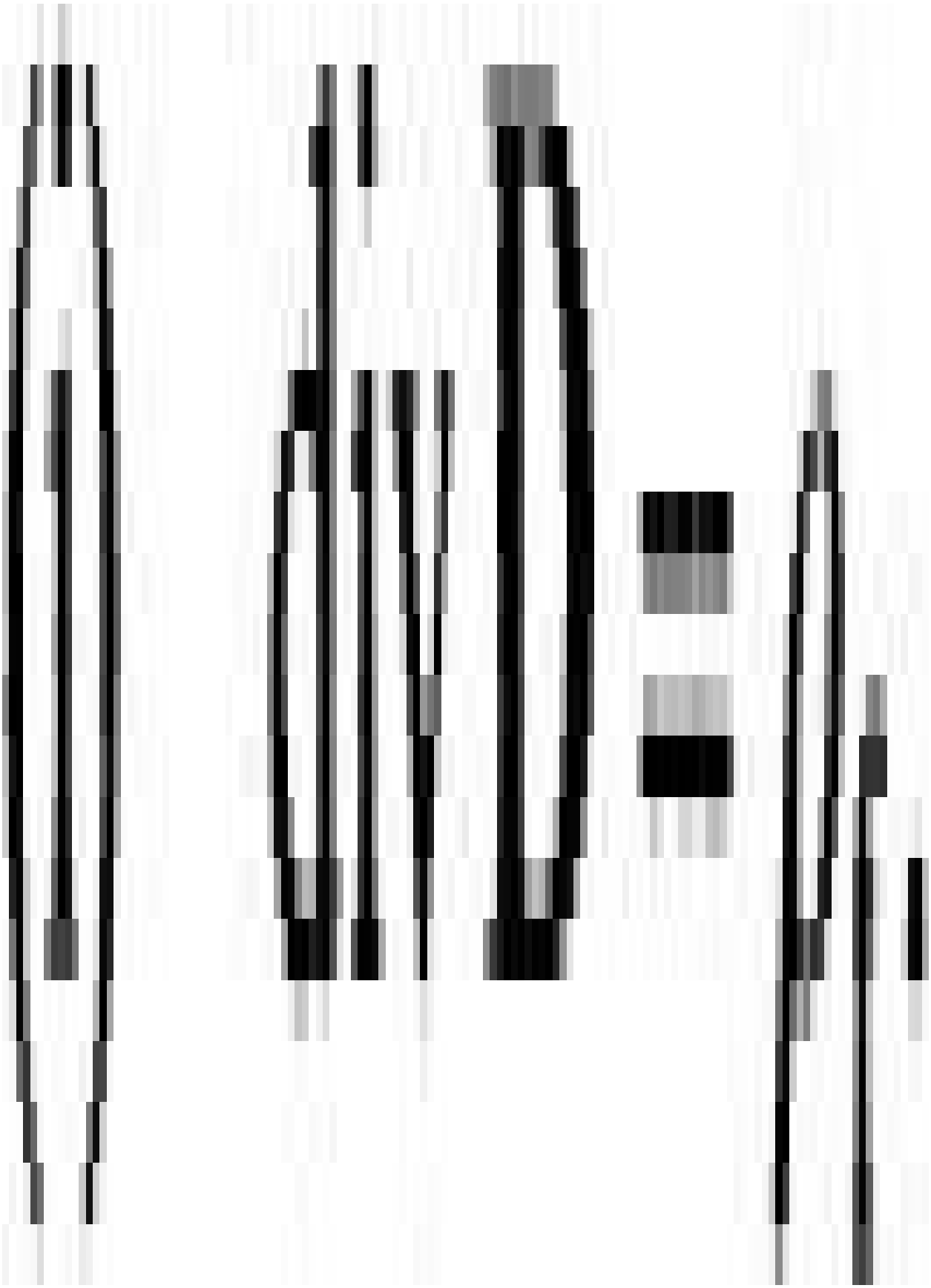
$$H = B / \mu \mu_0 ;$$

define the auxiliary vectors D and H in terms of the fields E and B and the quantities ε, ε0, μ and μ0. The relative permittivity ε and the relative permeability μ depend on frequency, as in principle do all parameters which describe the macroscopic behaviour of media in electric and magnetic fields. In addition we have discussed ferromagnetic materials for which the relationship between thefields B and H is complicated and not conveniently described in terms of relative permeability.

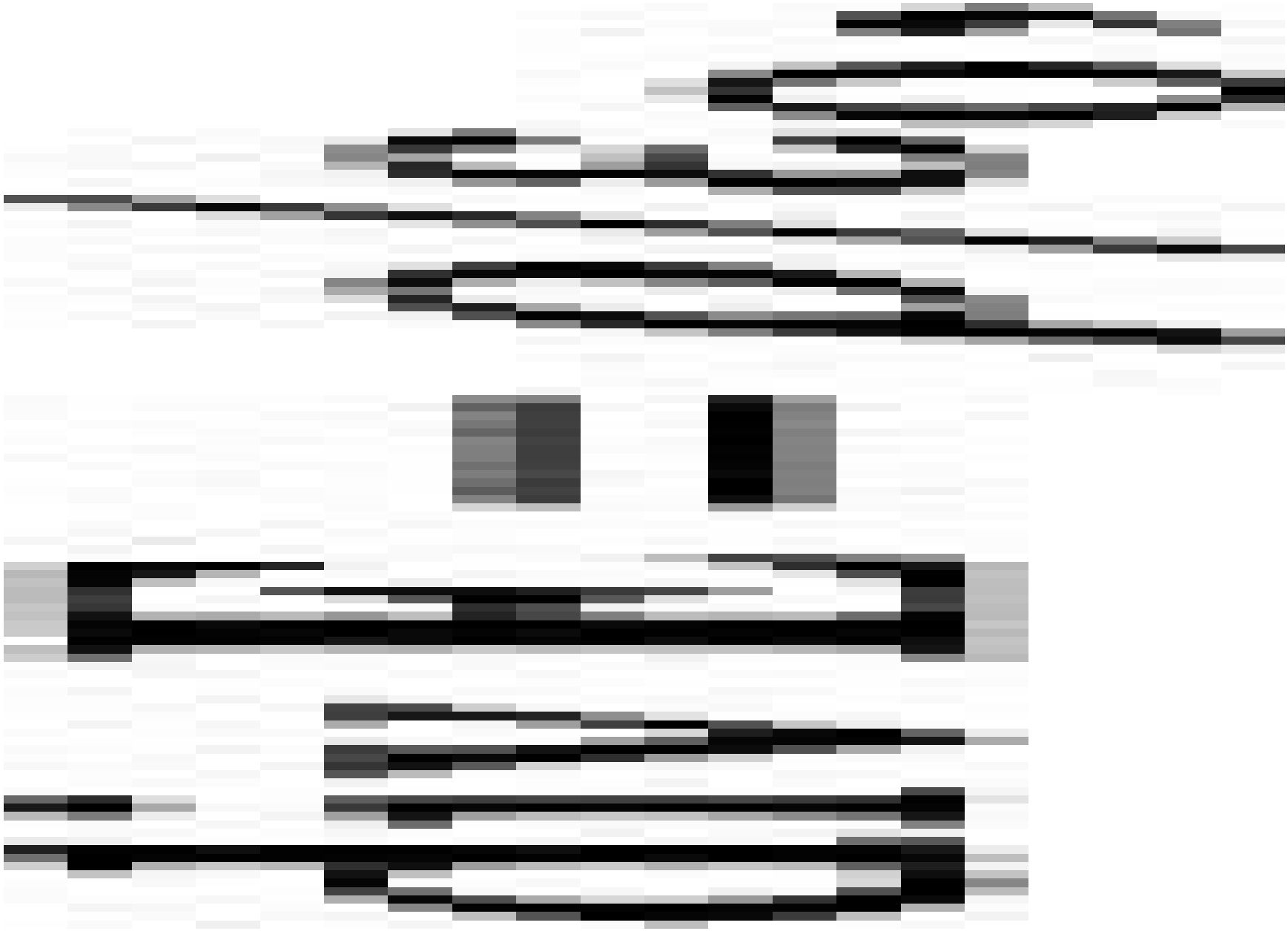
It may be noted that Equations (10.10) and (10.13) can be combined to give the equation of continuity (10.1). The law of conservation of charge is thus contained in Maxwell's equations, and is not an independent relationship between the free charge density ρf and the conduction current density jf. Maxwell's equations, together with the equation which gives the force on a charge q moving with velocity v in fields E and B, viz.



form a complete statement of the macroscopic theory of electromagnetism. Each equation is a generalization of a fairly simple experimental result and below we comment on the physical background to each one in turn.



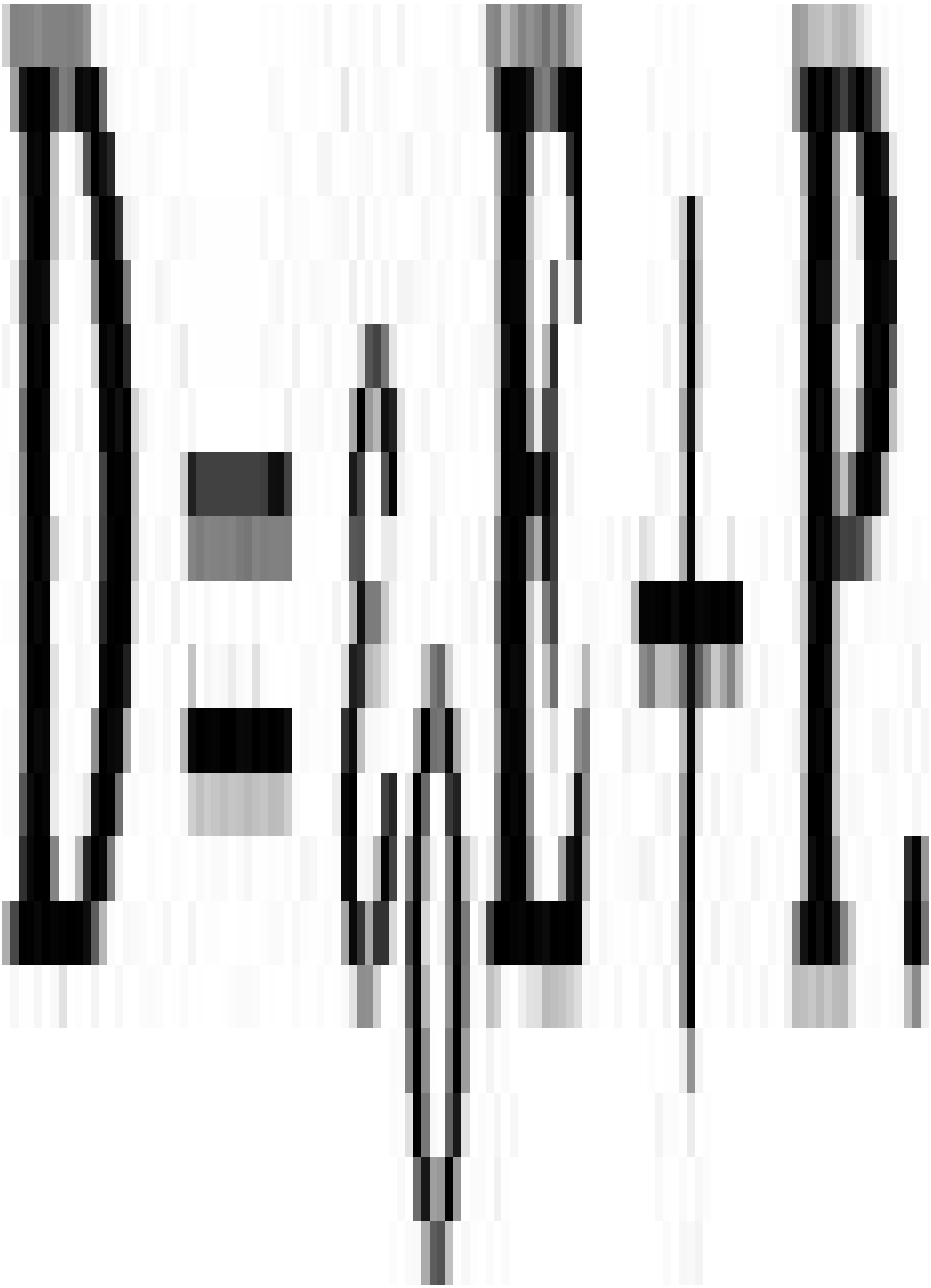
The inverse square law of force between stationary charges in free space leads to Gauss' theorem. In differential form this theorem states



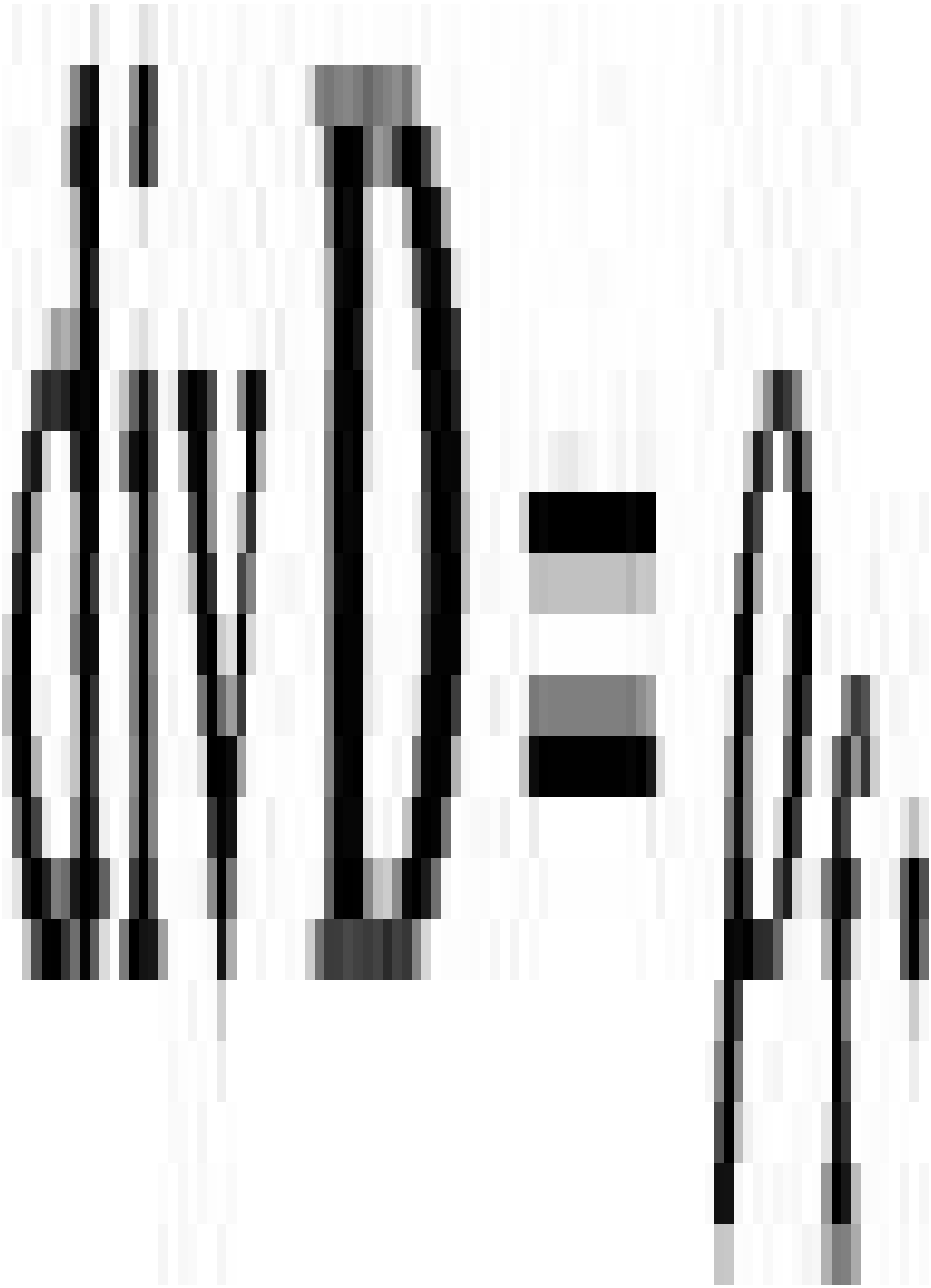
where ρ is the volume density of charge and ε0 is the permittivity of free space. If dielectric materials are present they acquire a polarization P and at each point are equivalent, for the purpose of calculating the electric field to which they give rise, to a volume density of charge (−div P). This volume density of charge must be added to the free charge density ρf, and we have

(10.14)

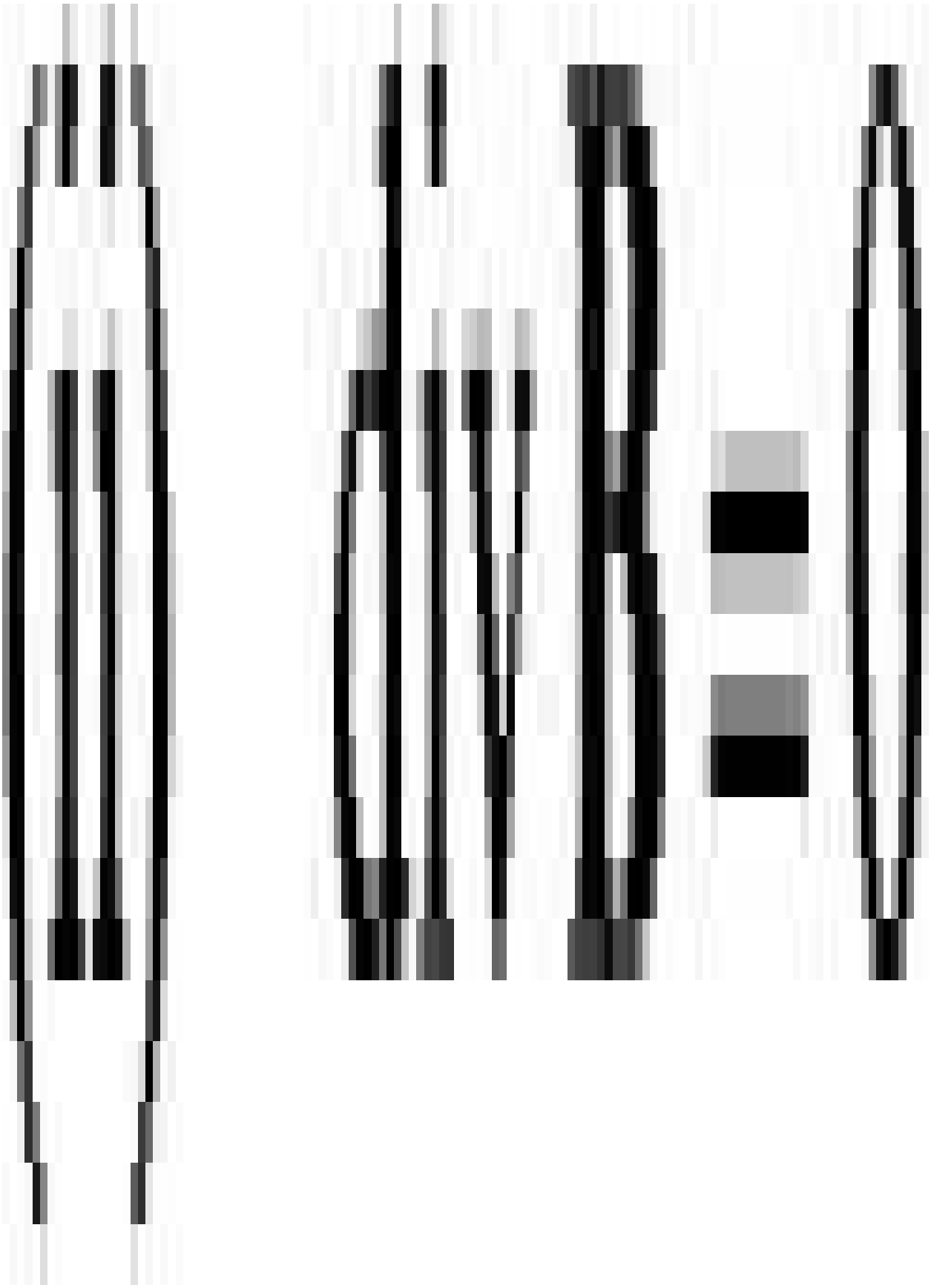
Rather than work with the polarization it is more convenient to use a vector D defined by



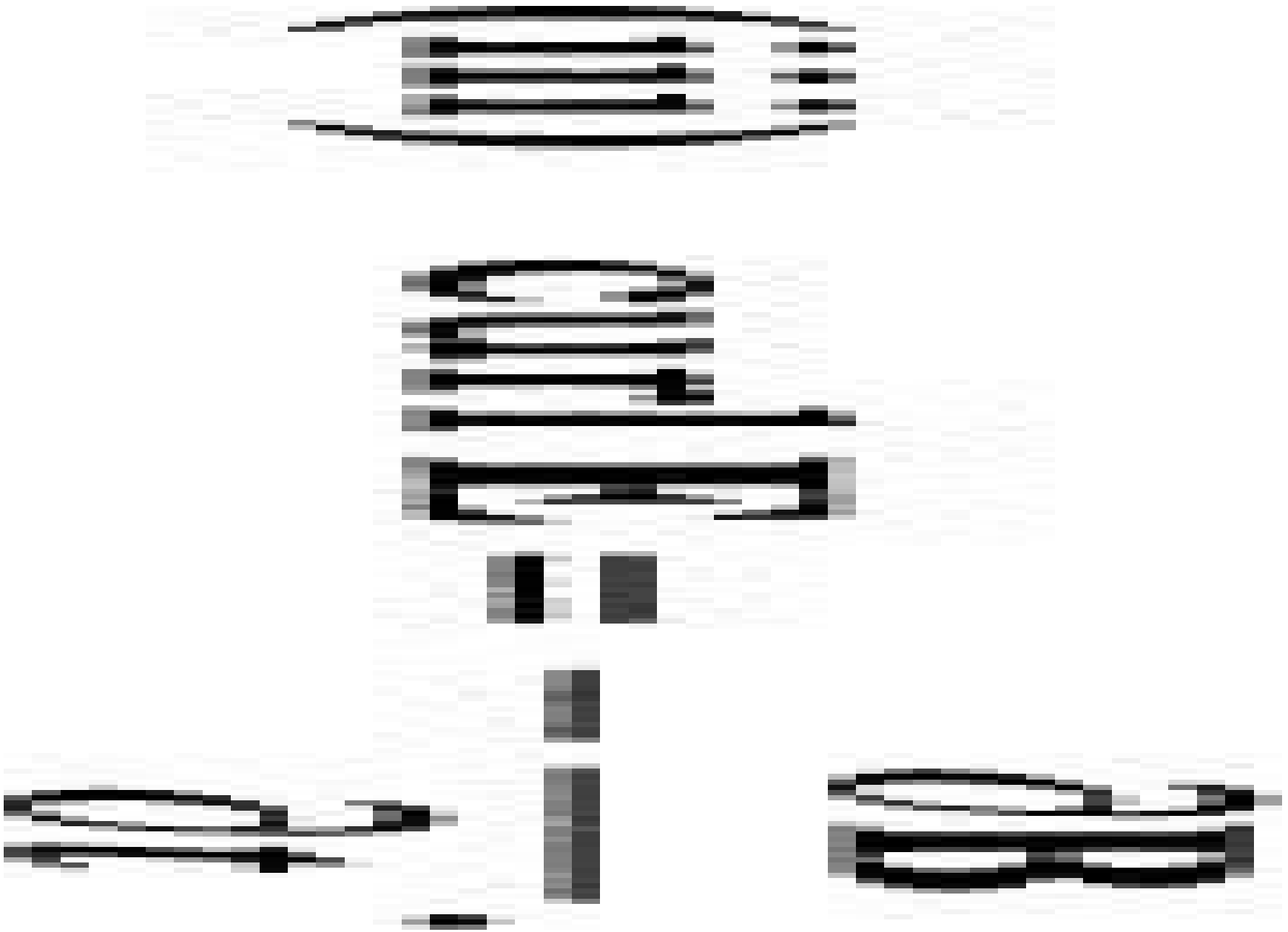
The electric displacement D is then related to the free charge density ρf by



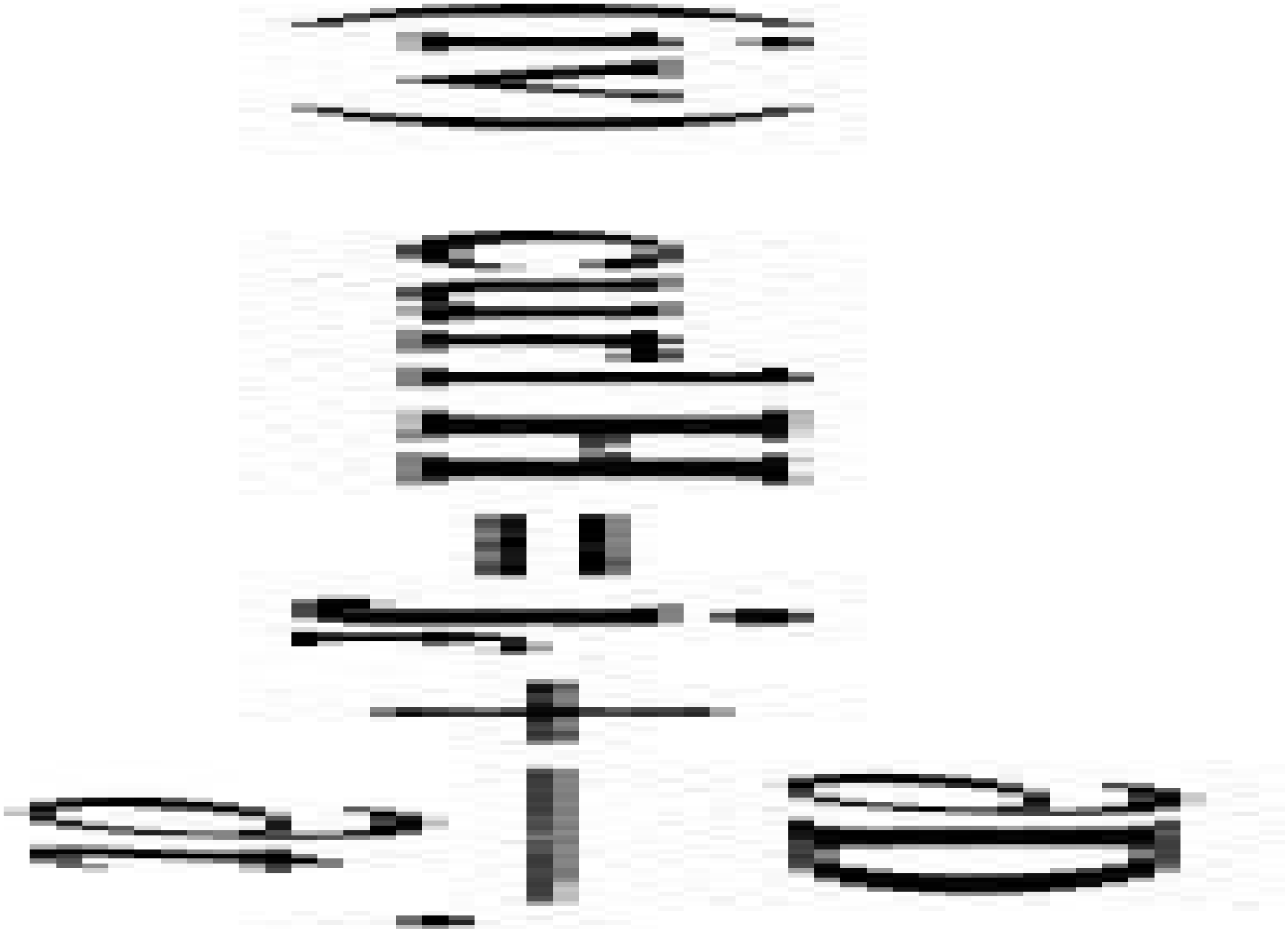
This equation, first discussed in electrostatics, is true for the total electric displacement from all sources, including electromagnetically induced fields and electric fields from moving charges. The divergence of the induced electric field is everywhere zero, and hence the field E in Equation (10.14) can be written as the sum of electrostatic and induced electric fields. Gauss' theorem is true for moving charges as well as for stationary charges, and Equation (10.14) also applies to the electric fields caused by moving charged bodies.



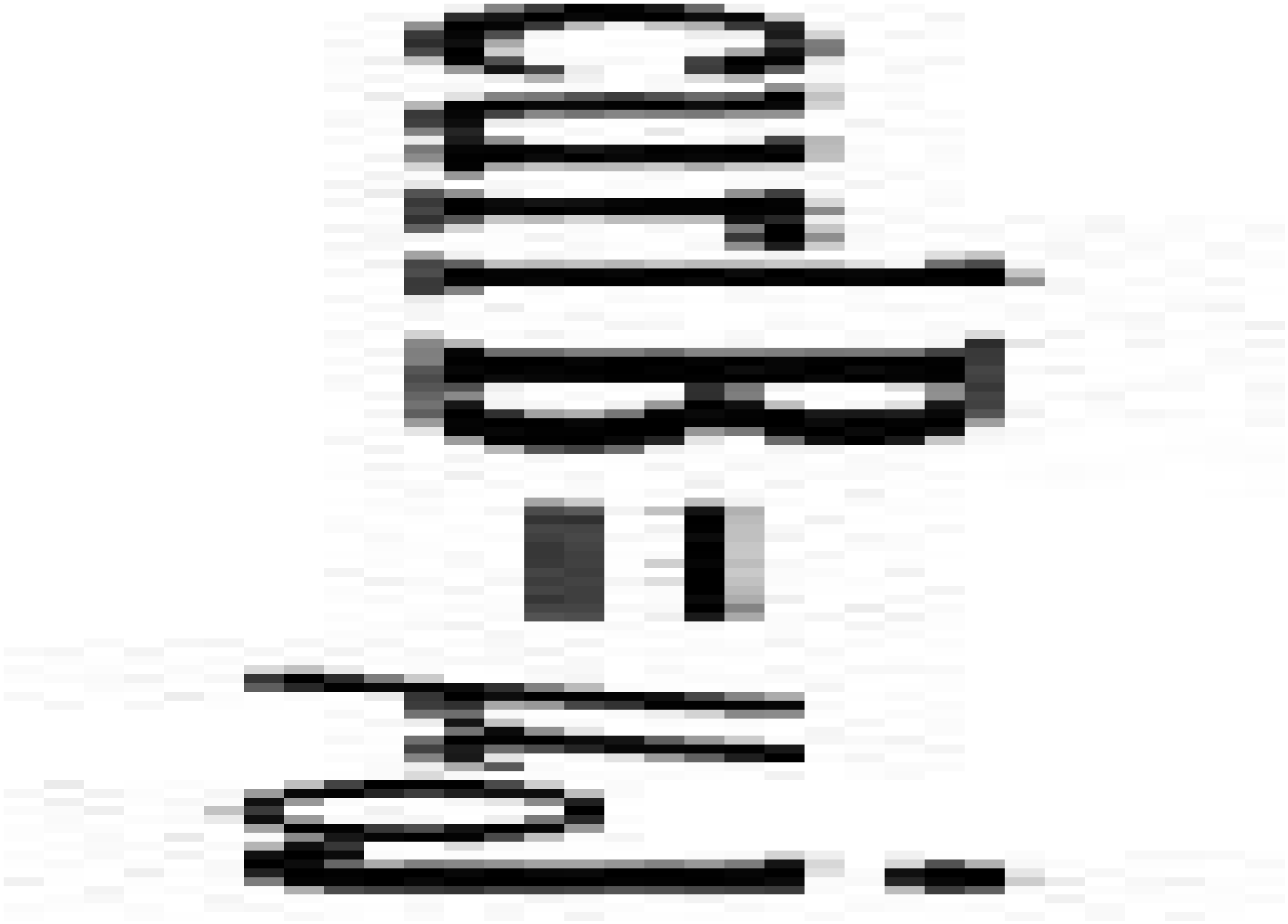
This equation is a consequence of the observation that all magnetic fields B have field lines that are continuous. There are no free magnetic ’poles’ or ’charges’.



Changing magnetic fields produce an electric field. This equation is the differential form of Faraday's law for the induced electric field. The integral form of the law expresses the experimental observation that the e.m.f. induced in a circuit is equal to the rate of change with time of the total magnetic flux through the circuit, no matter how the flux changes. The fields E and B are the total electric fields from all sources at each point.



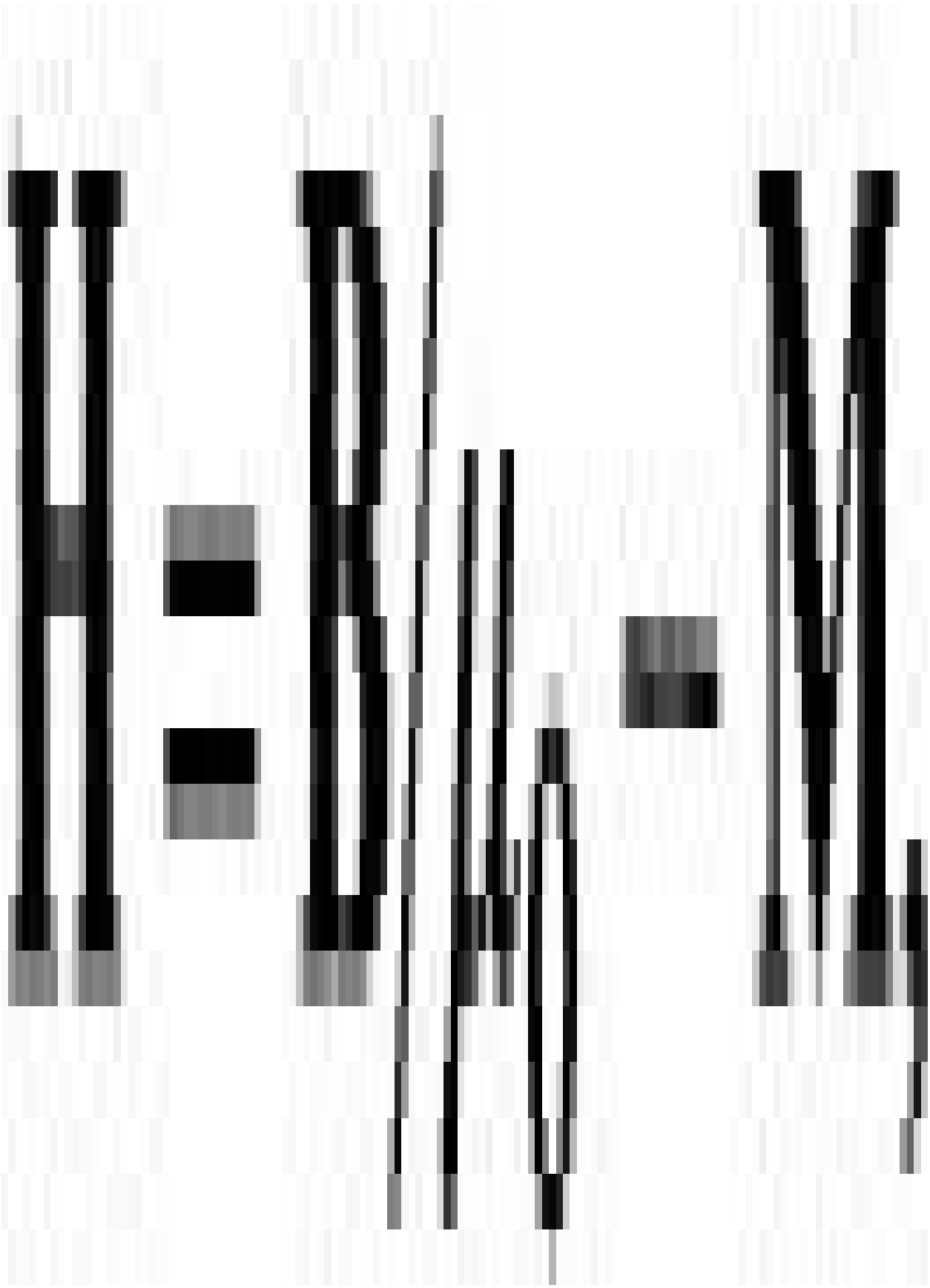
Ampère’s law for steady currents relates the line integral of the field B around a closed loop to the total current flowing through the loop. The law summarizes the results of experiments which show that two small coils carrying steady currents exert forces and torques on each other which show the same variation with separation, orientation, etc., as the forces and torques between two small electric dipoles. Ampère’s law in differential form relates the field B to the current density j at each point, and is



for currents flowing in free space. μ0 is the permeability of free space. If matter is present it acquires a magnetization M. At each point the magnetized material, for the purpose of calculating the magnetic field to which it gives rise, is equivalent to a current density curl M. This magnetization current density must be added to the free current density jf, and we have

(10.15)

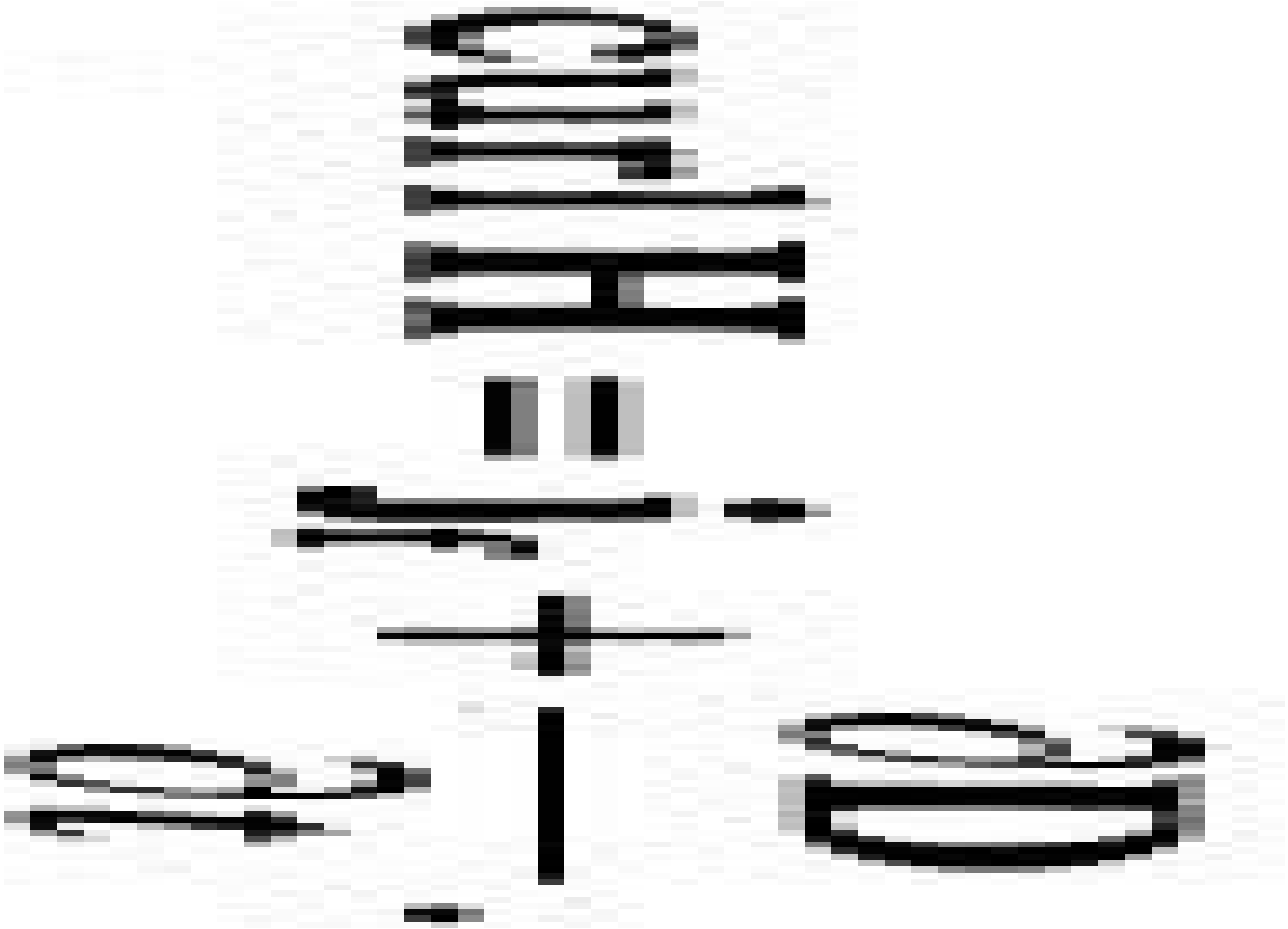
It is more useful to introduce the magnetic intensity H than to work with the magnetization. The magnetic intensity is defined by the equation



and, rearranging Equation (10.15), is related to the free current density by the equation

(10.16)

If the currents change with time the displacement current density has to be included, as discussed in section 10.2, and Equation (10.16) becomes

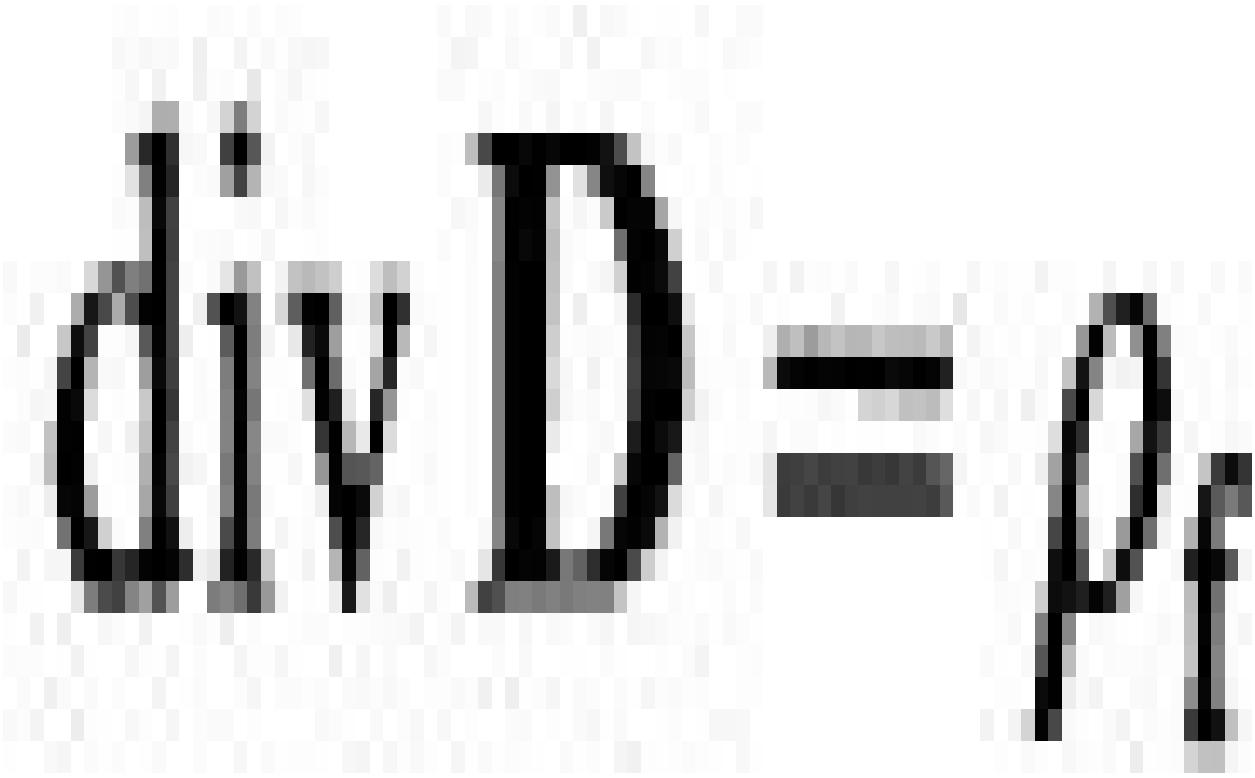


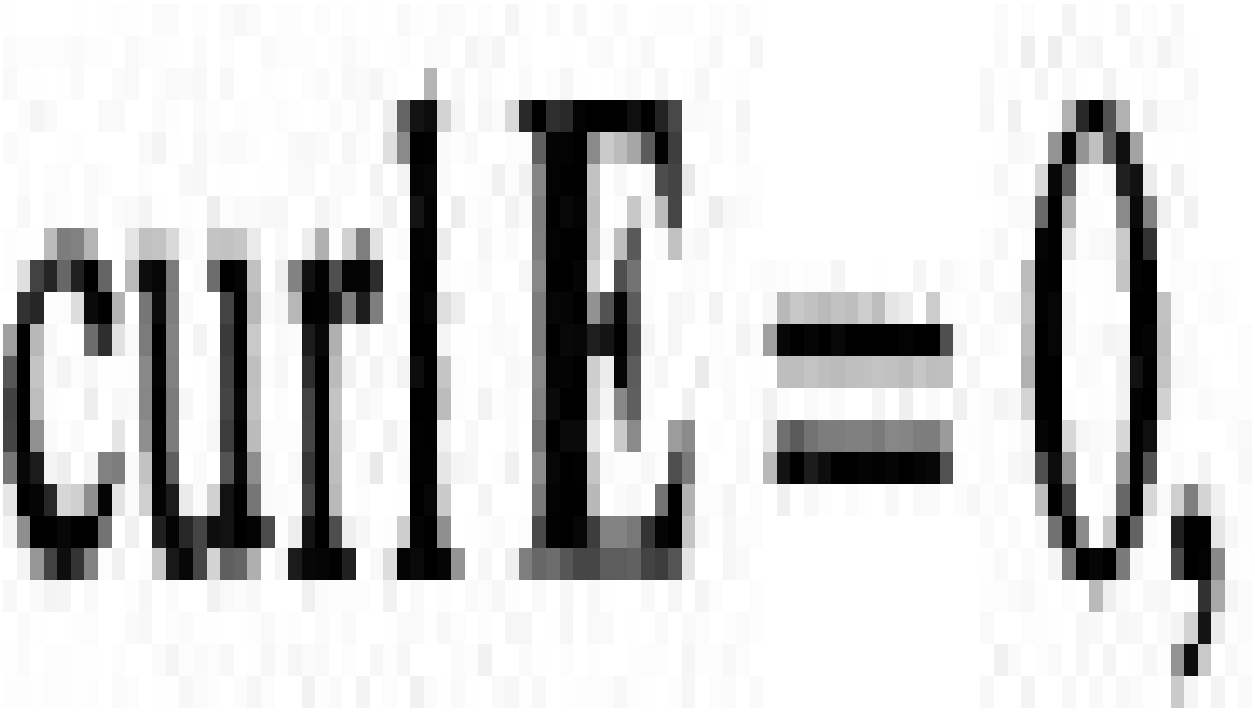
Much of the remainder of this book deals with mathematical solutions to Maxwell's equations in different situations, i.e. with different boundary conditions and source distributions. There is thus a lot of mathematics in the study of certain parts of electricity and magnetism. However, many important features in most situations can be understood by using physical insight to suggest approximations to simplify the mathematics.

The development of the macroscopic theory of electromagnetism—Maxwell's equations—beginning around the year 1785 with the work of Coulomb on the forces between charged bodies, and ending around 1867 with Maxwell, is an elegant example of the scientific method. The many practical applications of electromagnetic theory are a constant reminder of the benefits which may result from scientific curiosity.

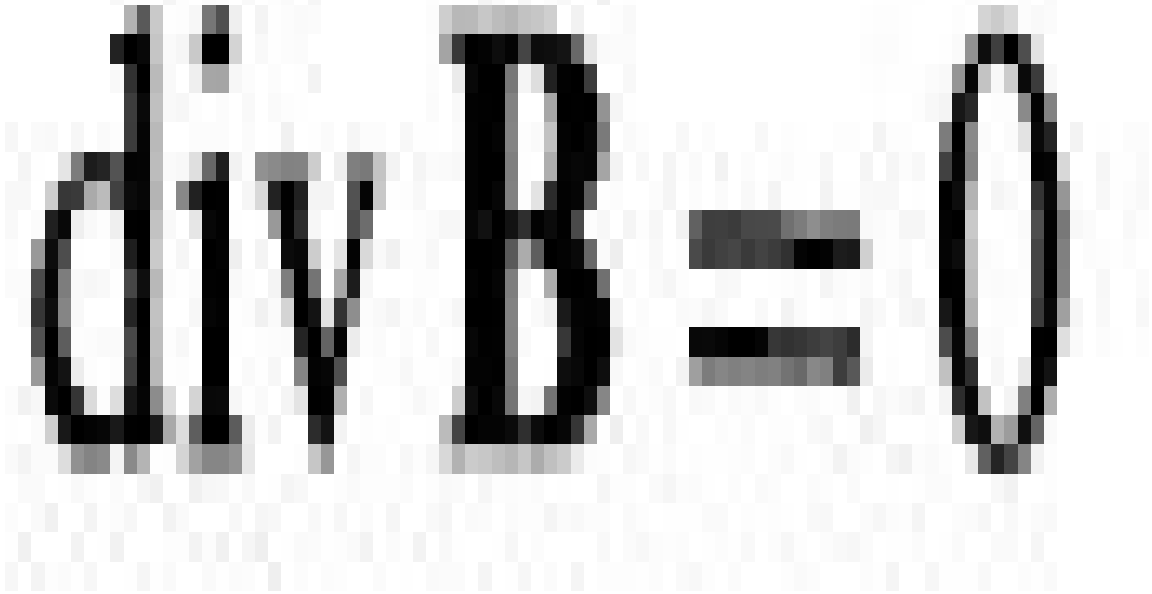
10.4 ELECTROMAGNETIC RADIATION

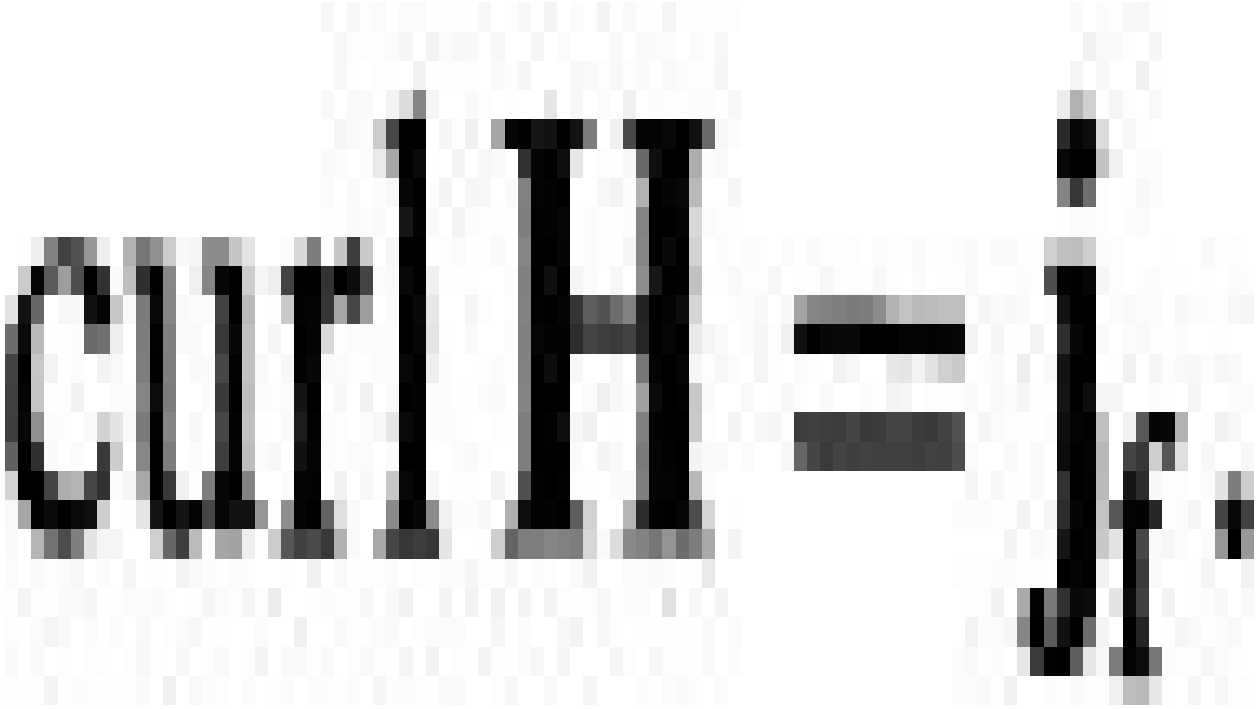
If the electric and magnetic fields do not vary with time Maxwell's equations reduce to the equations of electrostatics and magnetostatics discussed in Chapters 1 to 5. For steady fields Maxwell's equations become:


div D = ρf


curl E = 0,

and


div B = 0


curl H = j_f

If the fields vary slowly with time, so that the displacement current is negligible, we have what is often called a quasi-static situation, and the fields obey similar equations to those above, except that the second is replaced by Faraday’s law,

curlE = − ∂B/∂t

For fields which vary rapidly with time, typically if they have frequency components of 100 kHz or higher, the displacement current has to be included. This gives rise to new solutions to Maxwell's equations corresponding to electromagnetic radiation.

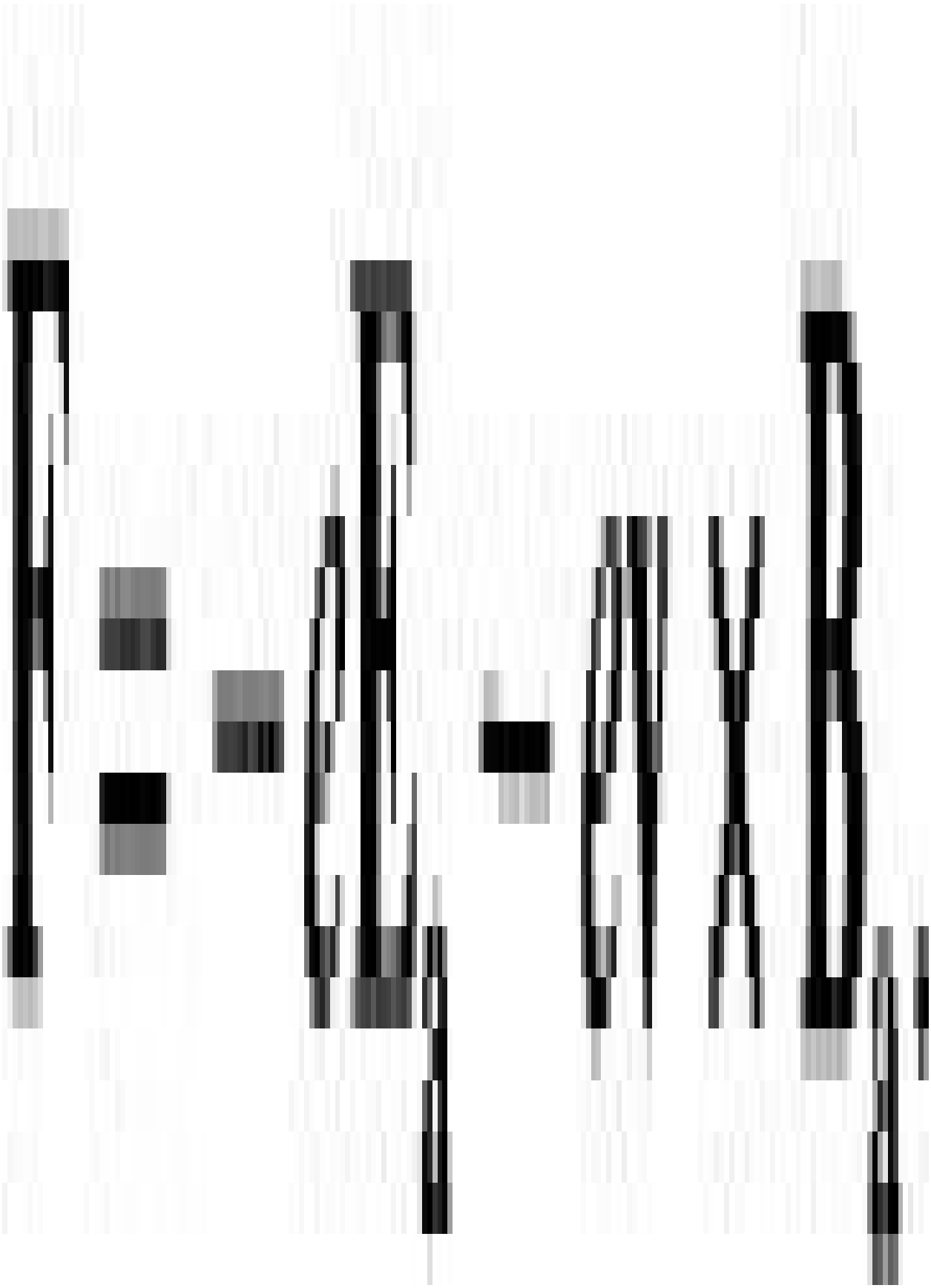
A simple example in which the fields change rapidly is provided by a spark breakdown between two electrically charged spheres. The large, but short-lived, current in the spark produces a changing magnetic field which induces an electric field near to the spheres. Prior to the addition of the displacement current term in Equation (10.13) one would have believed that the electric and magnetic fields would quickly die away to zero, as they do to all intents and purposes when a battery supplying a current through an inductive coil is suddenly removed. With the addition of displacement current however, the rapidly changing electric field can now in its turn induce a magnetic field. Solutions to Maxwell's equations are now possible in which the changing electric and magnetic fields combine in such a way as to produce new fields further out from the spheres as the old ones die away. This process is continually repeated, and the result is that a combination of related electric and magnetic fields travels outwards from the spheres with a speed which is determined by the electromagnetic field equations (10.10)–(10.13). The fields have energy, and some of the energy dissipated in the spark moves outwards from the spheres with the radiated fields. The spark is said to have radiated energy, which is carried away in the electromagnetic radiation field. In the next chapter we derive wavelike solutions to Maxwell's equations in free space that describe electromagnetic disturbances which are propagated at a fixed speed. In this section, however, we simply note that the quantity $(\varepsilon_0\mu_0)^{-1/2}$ has the dimensions of velocity, and that this is the speed of travel of the electromagnetic radiation field in free space. This speed is usually given the symbol c and its numerical value is very close to $3 \times 10^8 \mathrm{ms}^{-1}$.

Hertz in 1887 first detected electromagnetic radiation by inducing an electric field in a circuit some distance away from a second circuit in which there were rapidly changing currents. He showed that the speed of the radiated field was

roughly 3×10^8 ms−1, and also that the radiated field had many other properties similar to visible light. For example it could be reflected, refracted and polarized. The conclusion that visible light is an electromagnetic wave in which electric and magnetic fields vary rapidly in time was an early spectacular result of Maxwell’s proper formulation of the laws of electromagnetism.

10.5 THE MICROSCOPIC FIELD EQUATIONS

Throughout this book we have attempted to keep sight of the fact that the macroscopic electric and magnetic fields inside media are averages of microscopic fields. The microscopic fields are produced by charges on electrons and protons moving in free space and by the intrinsic magnetic dipole moments of electrons, protons and neutrons. In this section we will have expressions for the fields and parameters used in the macroscopic theory in terms of microscopic quantities. For convenience we will use the symbols Ea and Ba to denote the atomic fields which describe forces on microscopic charges. The force on an electron of charge −e, moving with velocity v is given by



It is possible to formulate the laws of electricity and magnetism in terms of the fields Ea and Ba alone, but it is easier on a macroscopic scale to use the macroscopic fields E, D, B and H.

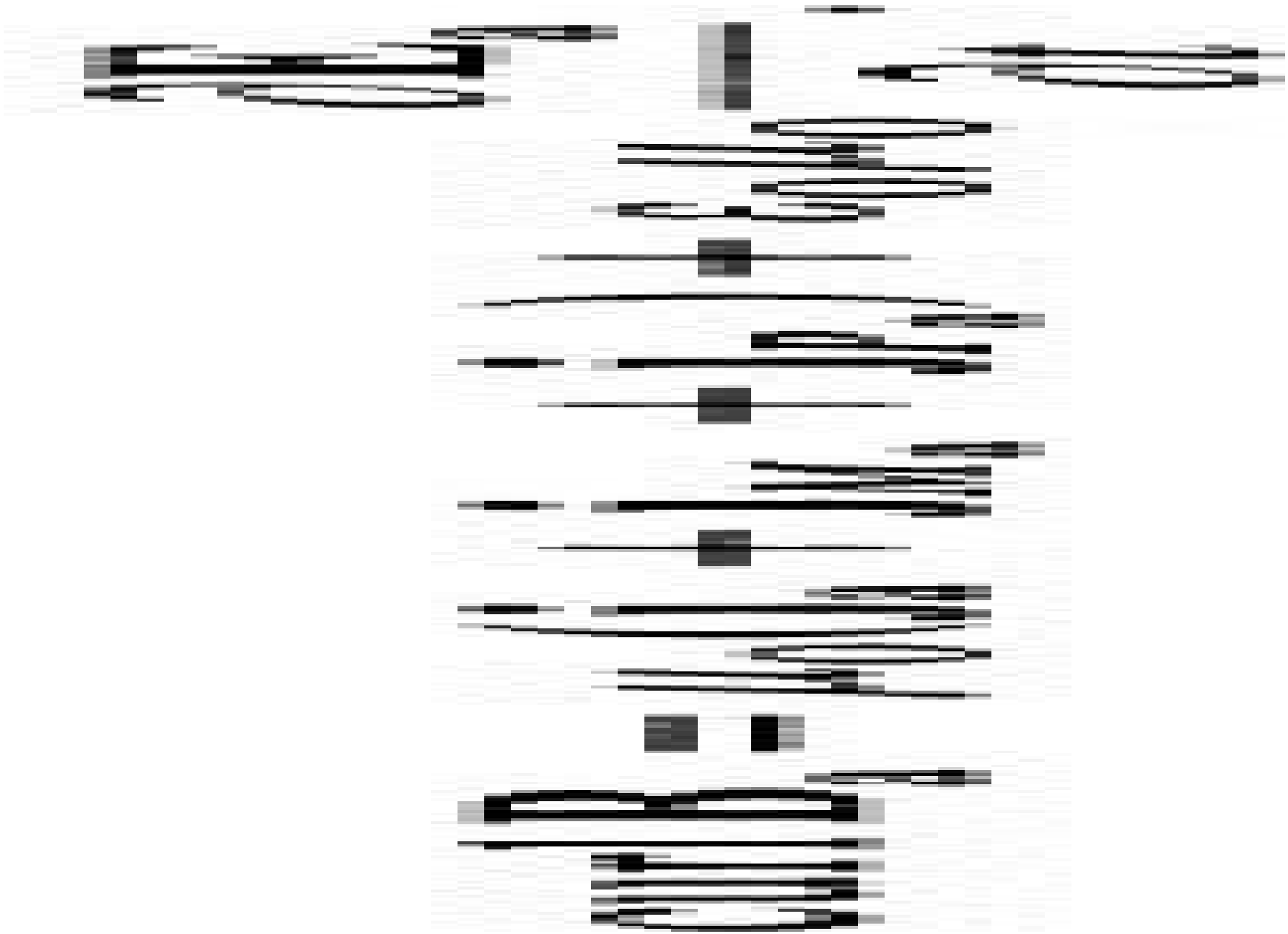
The atomic fields obey relationships similar to Maxwell's equations:

(10.17)

(10.18)

(10.19)

(10.20)



In these equations ρa is the microscopic volume density of charge and ja is the microscopic current density. If va is the average velocity of the charges constituting the charge density pa, the current density is given by is the current density equivalent to the microscopic magnetization Ma, which is due to the distribution of intrinsic magnetic dipole moments and is the current density which arises from the microscopic polarization Pa varying with time. Equations (10.17) to (10.20) are in a sense more fundamental than Maxwell's equations for the macroscopic fields, since the atomic field equations hold at every point, whereas Maxwell's equations involve averages over very many atoms. The macroscopic electric and magnetic fields are defined by averaging the microscopic fields over volumes very small on a macroscopic scale but which contain very many atoms. We write the macroscopic fields as

$$\vec{E} = \vec{E}_d$$

$$\vec{B} = \vec{B}_d.$$

Maxwell’s equations for the fields E and B can be written in terms of the polarization P and the magnetization M by substituting for the fields D and H in Equations (10.10) to (10.13), using Equations (2.20) and (5.27):

$$\operatorname{div} \mathbf{E} = (\rho_f - \operatorname{div} \mathbf{P})/\varepsilon_0$$

$$\operatorname{div} \mathbf{B} = 0$$

$$\operatorname{curl} \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$

$$\operatorname{curl} \mathbf{B} = \mu_0 \left(\mathbf{j}_f + \operatorname{curl} \mathbf{M} + \frac{\partial \mathbf{P}}{\partial t} \right) + \varepsilon_0 \mu_0 \frac{\partial \mathbf{E}}{\partial t}.$$

It can be proved that averaging the microscopic field equations over volumes which are very small on a macroscopic scale gives a set of four equations for the fields E and B which can be cast in a form similar to the above equations. The quantities P, M, ρf, and jf can then be identified with the following averages of microscopic quantities:

$$\mathbf{P} \equiv N\langle \mathbf{p} \rangle$$

$$\mathbf{M} \equiv N\langle \mathbf{m} \rangle$$

$$\rho_f \equiv -N_c e + N q_m$$

$$\mathbf{j}_f \equiv -N_c e \bar{\mathbf{v}}_c .$$

In these equations N is the number of atoms or molecules of the substance per unit volume, and Nc the number of free electrons per unit volume with average velocity , i.e. is their drift velocity as used in section 4.1. qm is the average charge on each atom or molecule; ⟨p⟩ is the average electric dipole moment of the molecules in the direction of the field E, and ⟨m⟩ the average magnetic dipole moment of the molecules* in the direction of the field B. The above definitions of P, M, ρf and jf are the same as the ones introduced in the earlier chapters when these quantities were first discussed. They enable us to define the auxiliary vectors D and H in terms of averages of microscopic quantities in the following way:

$$\mathbf{D} = \varepsilon_0 \bar{\bar{\mathbf{E}}}_a + N\langle \mathbf{p} \rangle$$

$$\mathbf{H} = \frac{\bar{\bar{\mathbf{B}}}_a}{\mu_0} - N\langle \mathbf{m} \rangle .$$

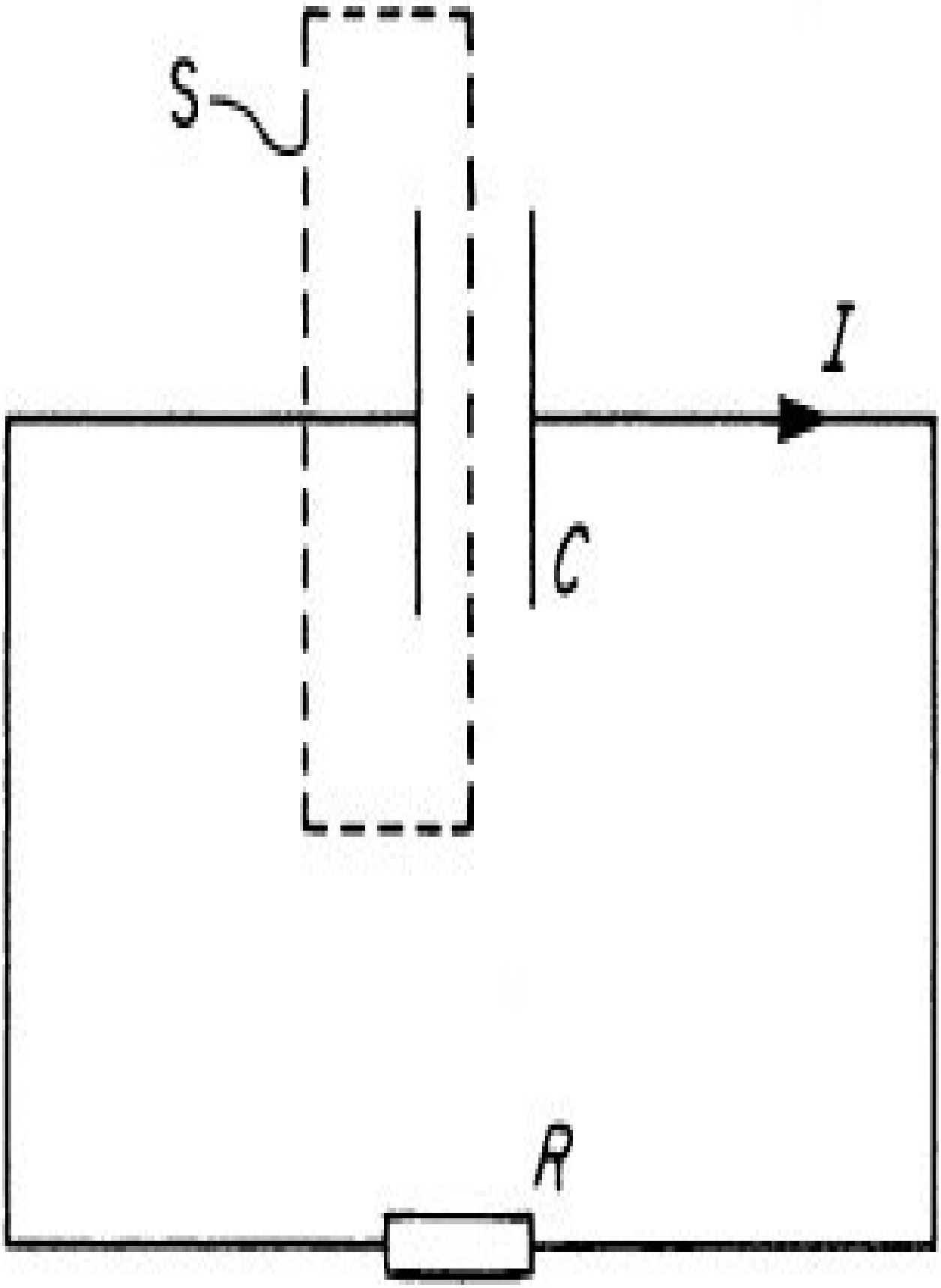
This concludes our correlation between macroscopic and microscopic quantities. The microscopic magnetization current density jMa is not always included in Equation (10.20). The intrinsic dipole moments may be considered to be due to the motion of microscopic charges within the particle itself, e.g. due to the electron spinning about an internal axis. In this case the intrinsic moments are considered to be included in the term ja. We follow the procedure outlined above because it leads naturally to an expression for the magnetic dipole moment of an atom or molecule (see Problem 10.3) which is taken over directly and used in quantum mechanics.

PROBLEMS 10

10.1 Positive charge leaves one plate of a parallel plate capacitor which is discharging through a resistor. At a certain time the rate of discharge is I amps. Calculate the displacement current flowing at this instant through a surface S which encloses one plate of the capacitor (see Figure 10.4). Show that the magnitude of the displacement current is equal to the conduction current I.

10.2 Suppose that a volume distribution of free charge density exists within a metallic conductor and is specified by the function $\rho(r, t)$. This function rapidly becomes zero as the charge redistributes itself over the surface of the conductor. If the direct current electrical conductivity of the metal is σ and it has N free electrons per unit volume, determine how $\rho(r, t)$ varies with time. How long does it take for the charge density at a certain point to drop to one half of its initial value in the case of copper? (The changes with time are so rapid that it is no longer a good enough approximation to use a simple linear relation between current density j and field E).

Figure 10.4.


S
I
C
R

10.3 Obtain expressions for the electric and magnetic dipole moments of an atom or molecule in terms of the microscopic quantities ρa, va and Ma defined in section 10.5. If the molecule is considered to consist of point particles obtain expressions for the moments in terms of the charges, velocities and intrinsic magnetic dipole moments of the point particles.

*The quotation is taken from An Introduction to the Meaning and Structure of Physics, by Leon N. Cooper. Harper and Row, New York, 1968.

*The average magnetic dipole moment of the molecules is here taken to include a contribution from the free electrons if there are any.

CHAPTER 11

Electromagnetic waves

We have already outlined in section 10.4 how a spark discharge can produce changing electric and magnetic fields that propagate outwards from the changing charge and current distributions in the spark. Any charge distribution which changes with time, or any varying current, can give rise to a radiated electromagnetic field. However it is only when the changes are rapid that the radiated fields are large enough to carry away an appreciable fraction of the power dissipated in the system. In practice radiated fields associated with varying currents are usually unimportant unless the variations contain frequencies higher than 100 kHz. This is about the lowest frequency used for commercial transmission of radio waves.

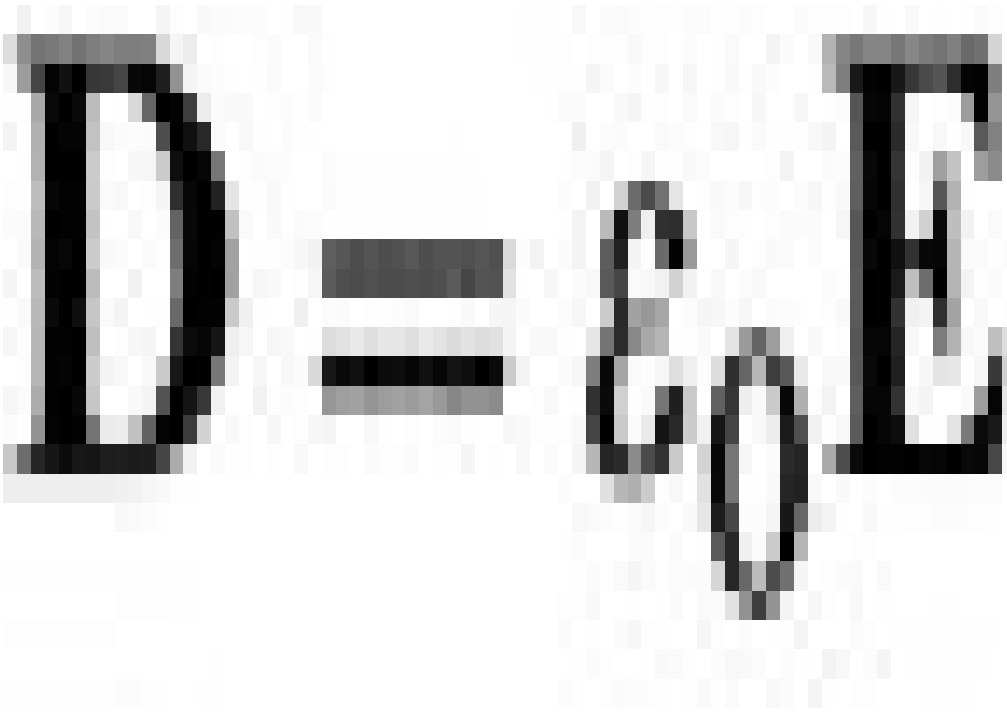
In this chapter we will consider some general properties of electromagnetic waves propagating in free space or in homogeneous isotropic media. We will also consider what happens to simple waves at a plane boundary between two media. The electromagnetic field will be treated classically. This means that we will assume the energy in the waves to be distributed continuously throughout the space occupied by them, and the energy to depend in some way on the macroscopic fields E, D, B and H. The same approach was adopted in the discussion of the energies in static electric and magnetic fields. The media in which the waves are travelling will also be treated classically, as if they were continuous, and will be characterized by macroscopic quantities such as relative permittivity ε, relative permeability μ, and electrical conductivity σ. These quantities vary with frequency, although the electrical conductivity of almost all substances is effectively constant for most frequencies that we will consider. The interpretation of the parameters ε, μ and σ throws much light on the microscopic description of matter, but these problems will rarely be mentioned in this chapter. The classical theory provides a good description of most electromagnetic

phenomena for frequencies below about 10^{11} Hz, and for some phenomena at higher frequencies.

11.1 ELECTROMAGNETIC WAVES IN FREE SPACE

In this section we will show formally that there exist solutions to Maxwell's equations in free space corresponding to waves which always travel with the same speed c.

Any electromagnetic field must satisfy all of Maxwell's equations. Let us assume that there is a changing charge and current distribution within a certain region of space, and let us consider the fields produced by this source of radiation. In the free space outside the region, the charge and current densities are everywhere zero, and the following relations hold between the vectors E and D, and between B and H:



Maxwell’s Equations (10.10) to (10.13) reduce to

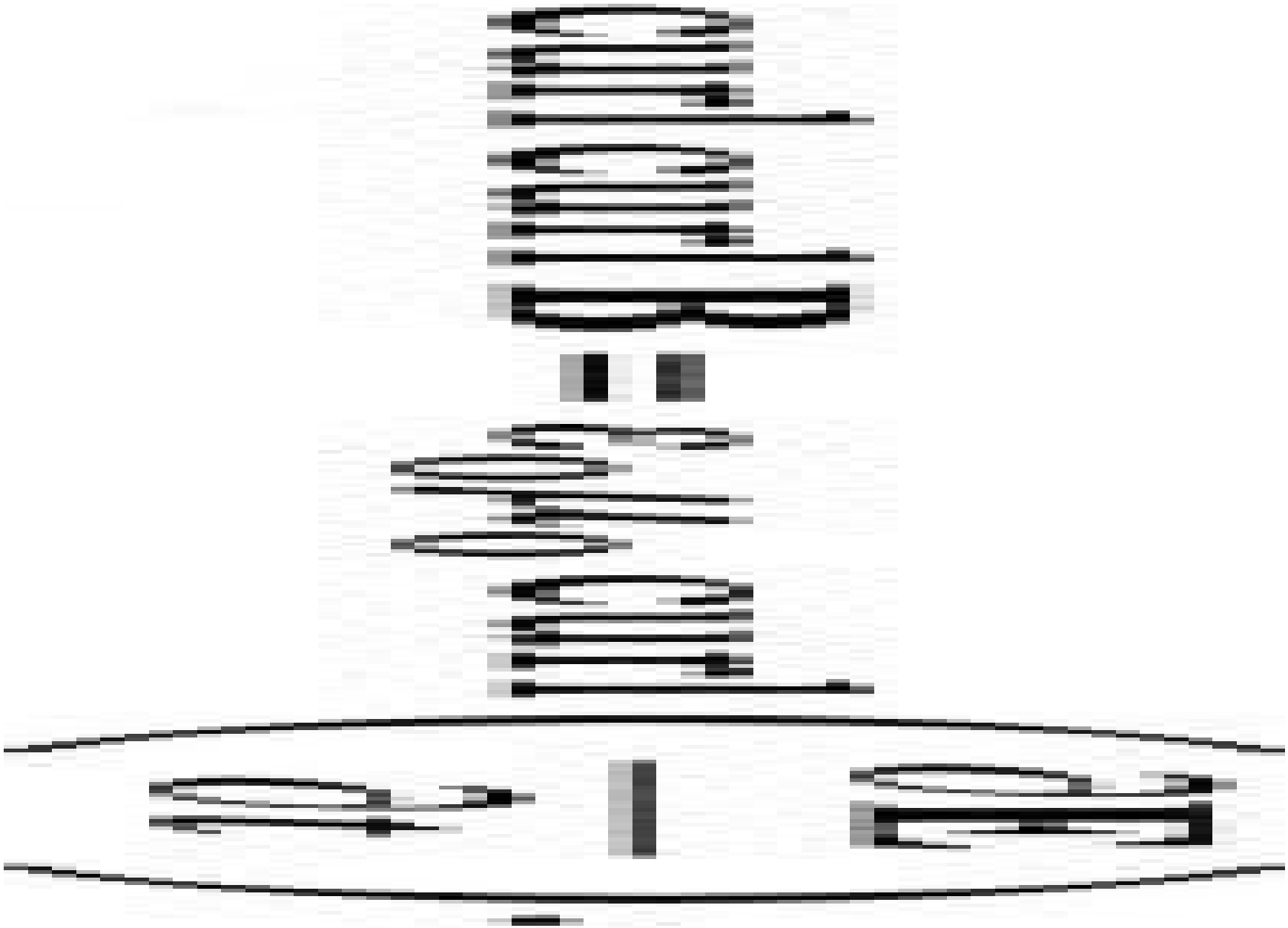
(11.1)

(11.2)

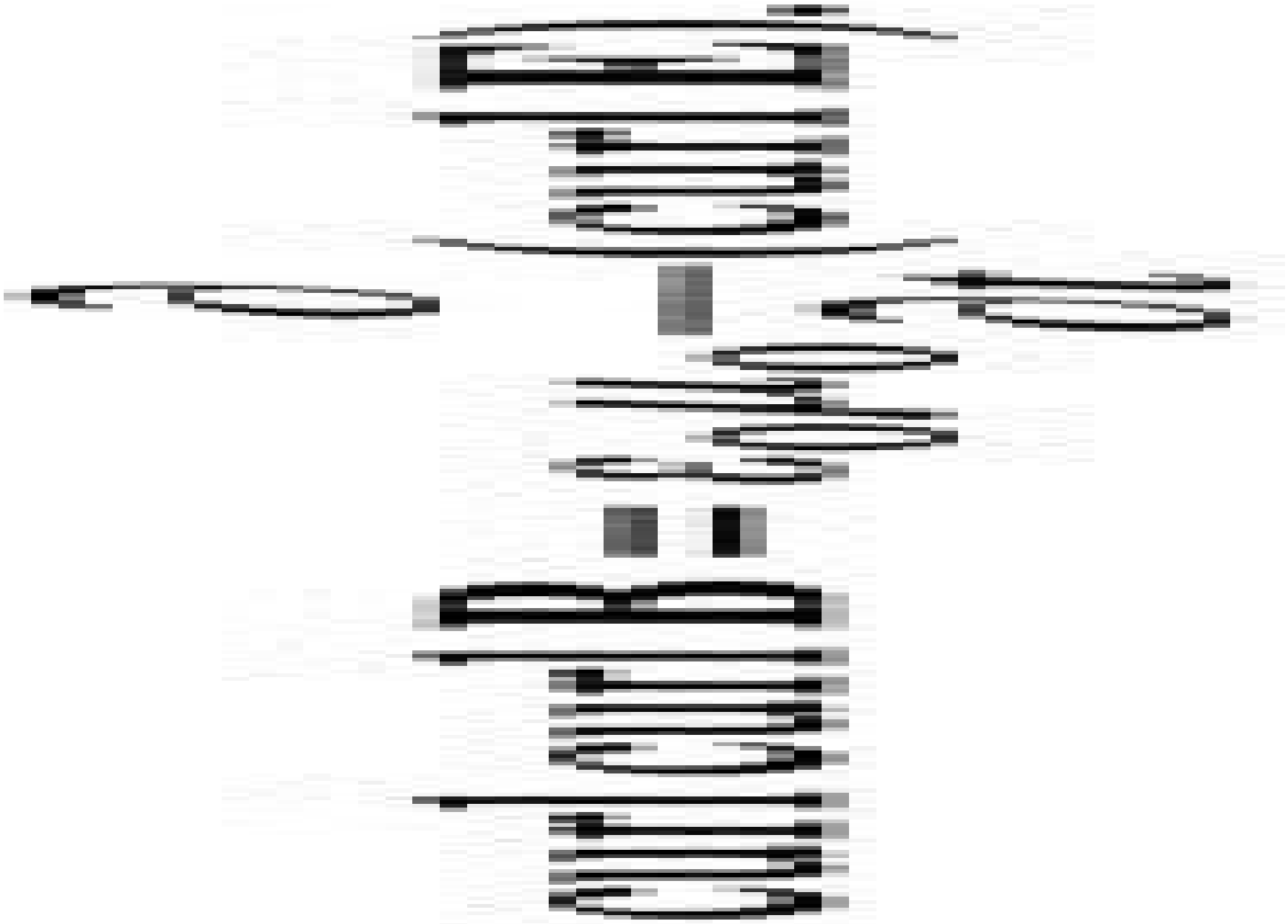
(11.3)

(11.4)

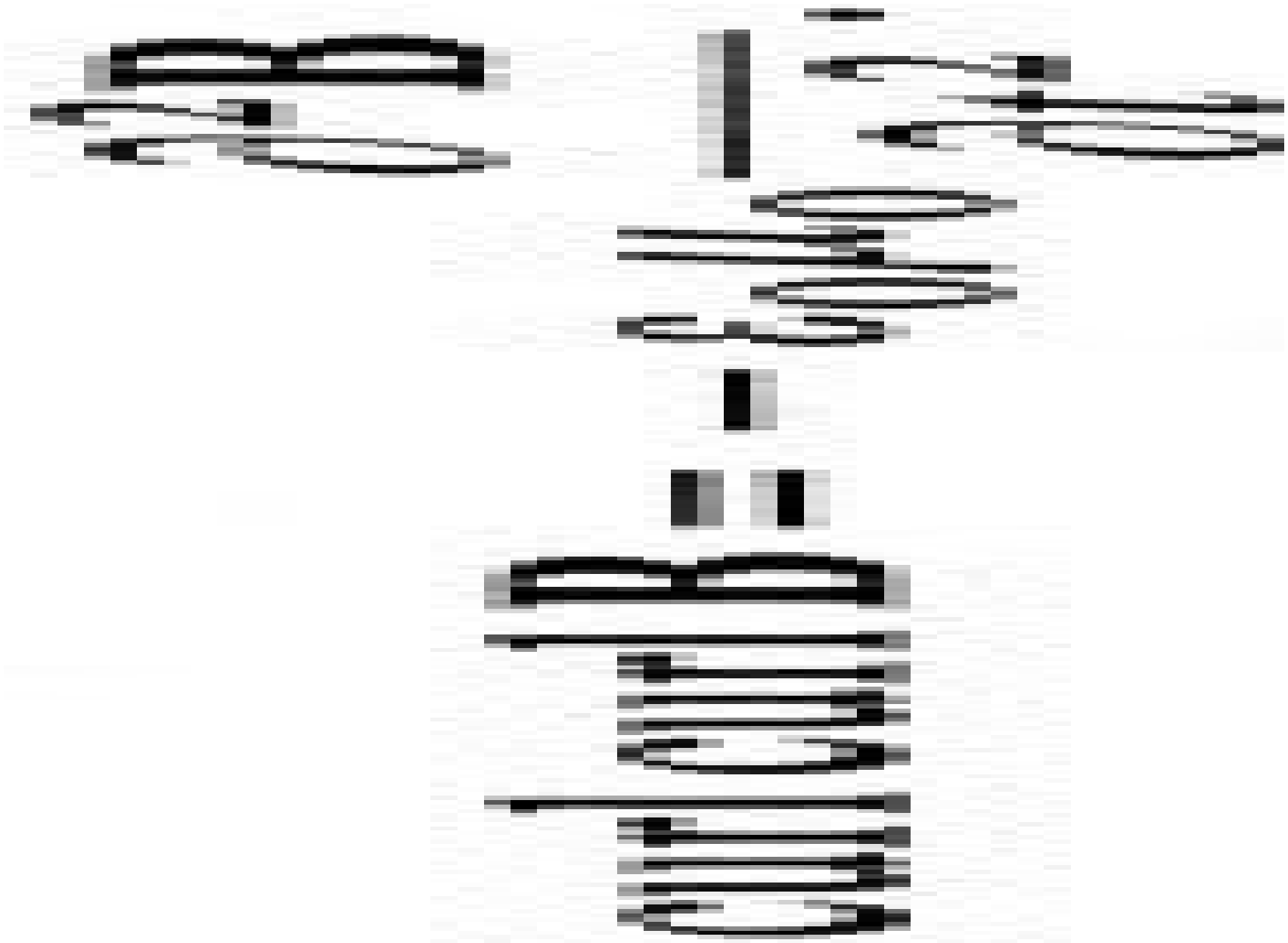
We will manipulate Equations (11.3) and (11.4) in order to derive equations obeyed by the fields E and B separately. Taking the curl of both sides of Equation (11.3) we find



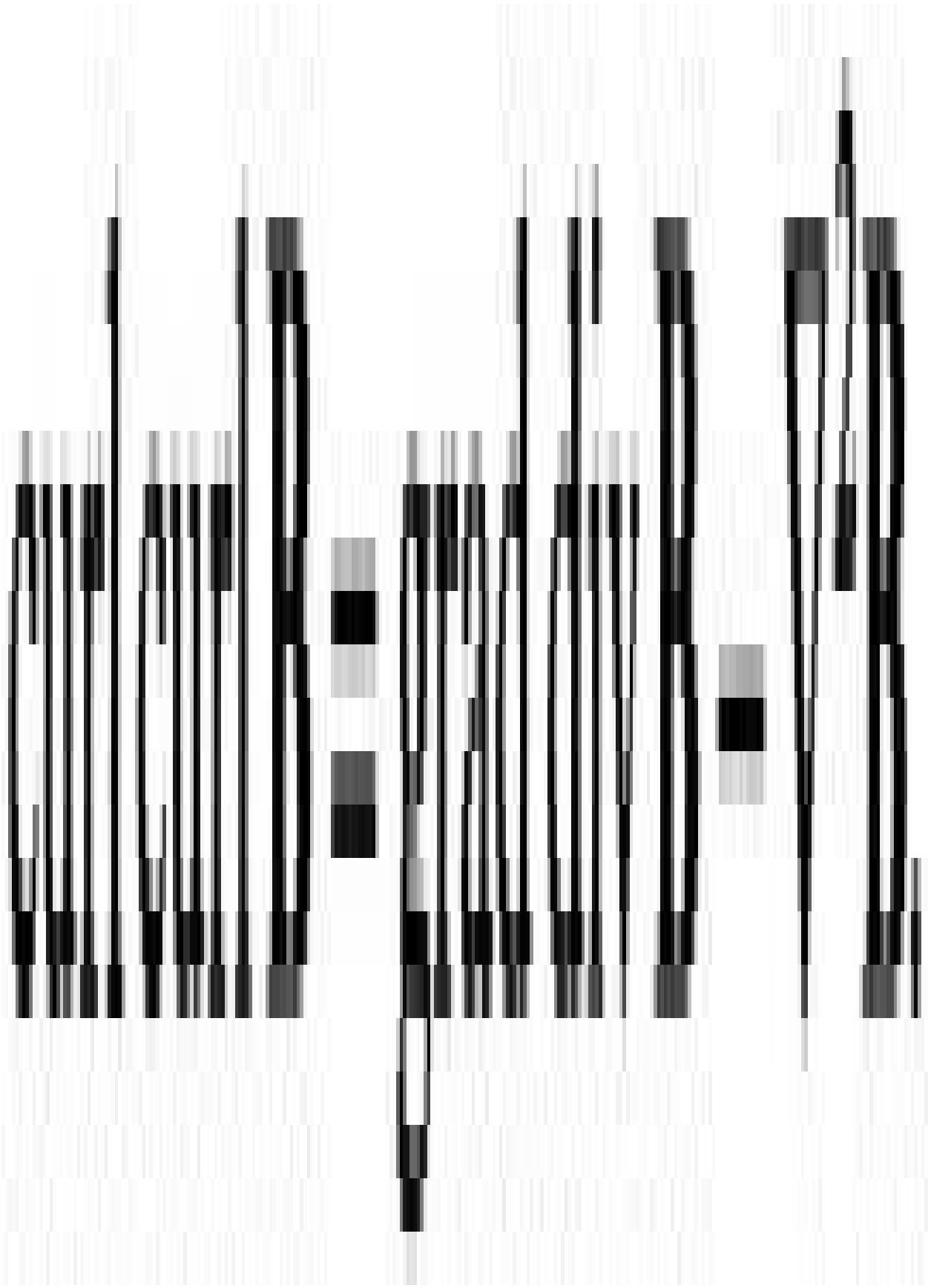
Since the operator curl does not include the time variable, curl ($\partial E/\partial t$) is the same function as ∂ (curl E)/∂t, and



Hence from Equation (11.4)



The function curl curl B can be replaced using the identity (B.21) given in Appendix B,



The divergence of the field B is everywhere zero and so we find finally

(11.5)

The magnetic field B obeys this equation everywhere in free space where the charge and current densities are zero.

By a procedure similar to that used above it can be shown that the electric field E must satisfy the equation

(11.6)

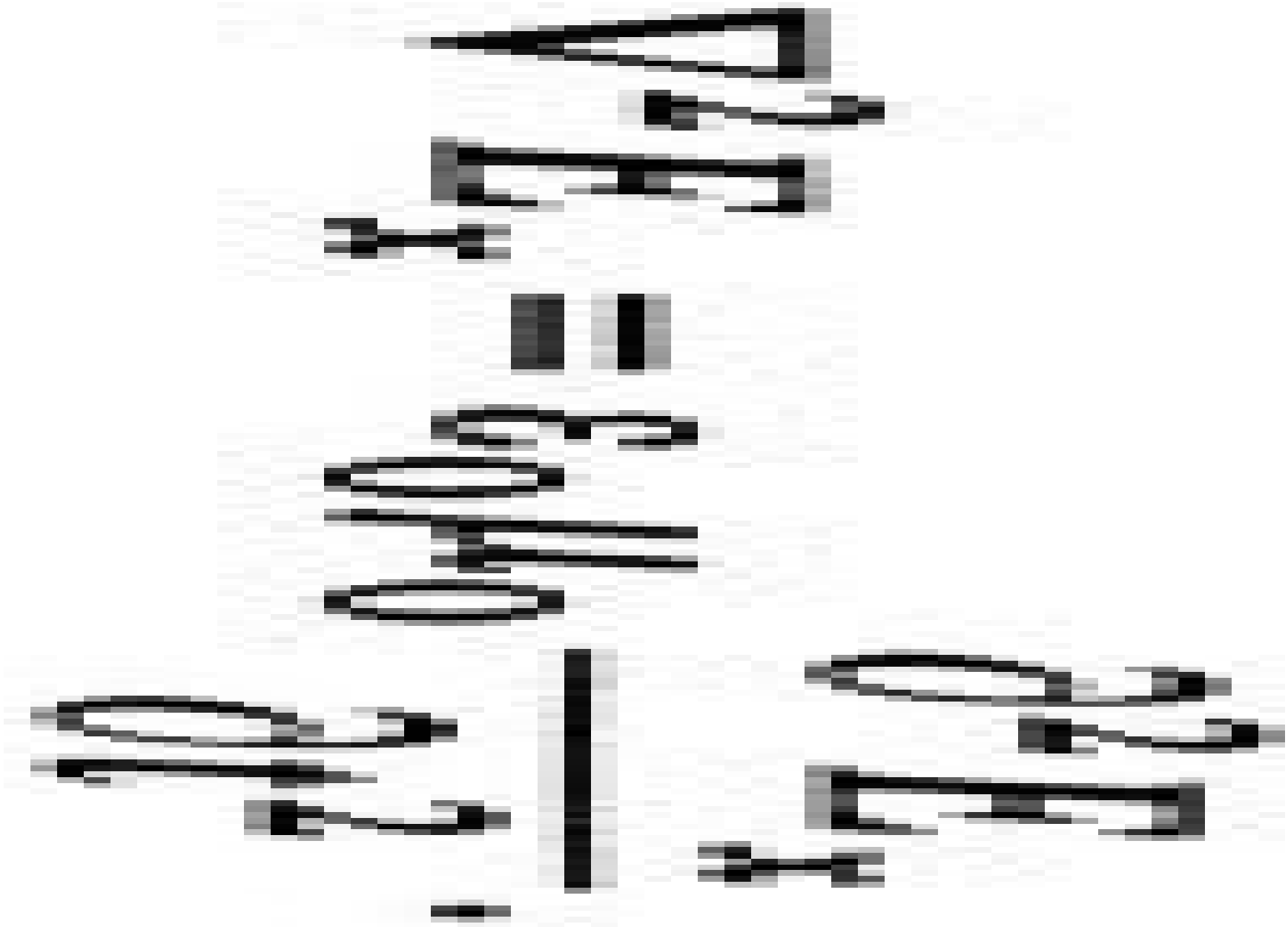
at all points in free space where there is no charge and current. Equations (11.5) and (11.6) are called the wave equations for the fields B and E.

The solutions to the wave equations are related electric and magnetic fields which together constitute an electromagnetic wave. It may be noted that the possibility of electromagnetic waves arises because of the introduction of displacement current. Without this current the right-hand side of Equation (11.3) would be zero and the set of Equations (11.1) to (11.4) would not have reduced to Equations (11.5) and (11.6) which describe waves.

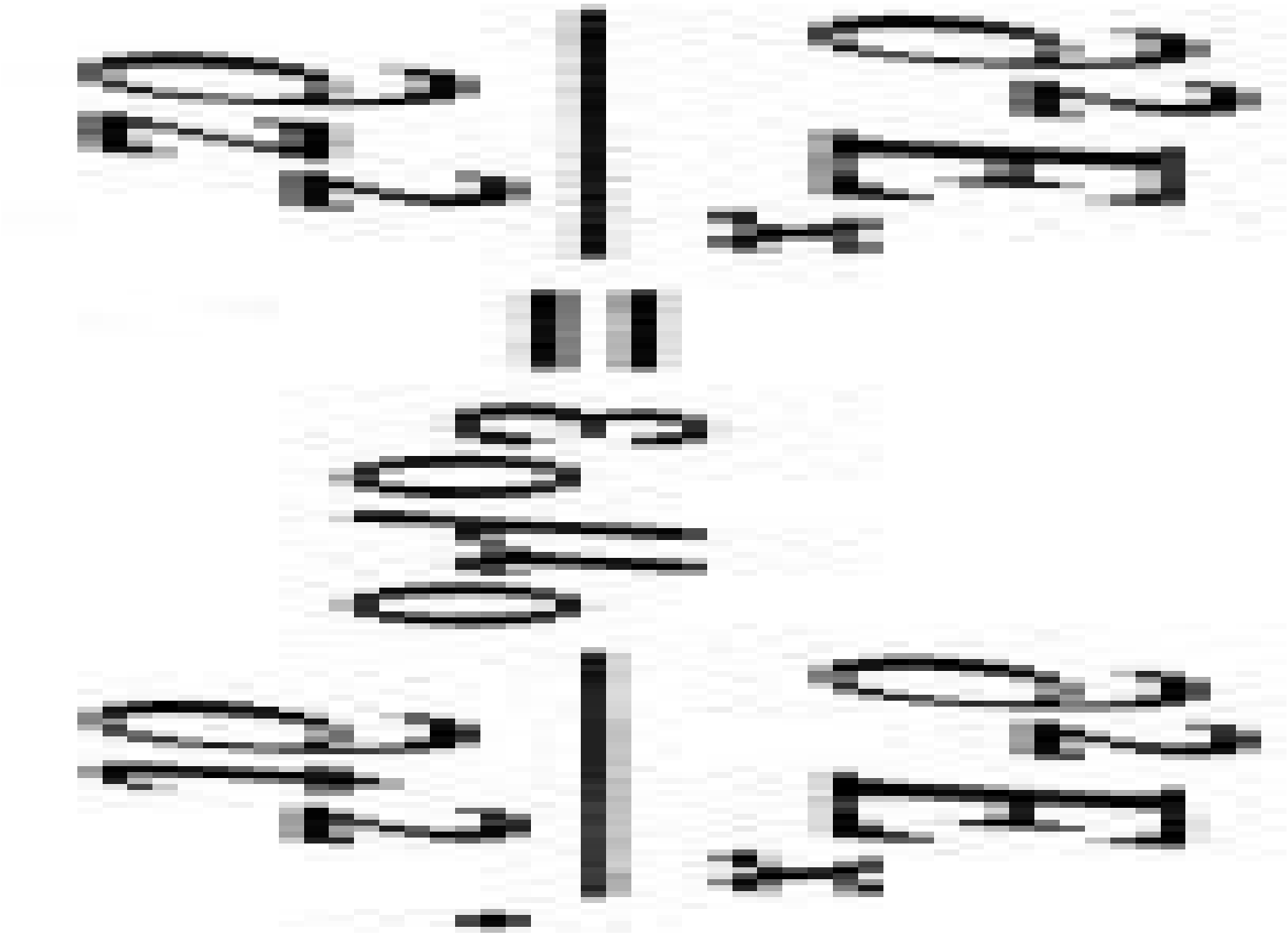
The wave equations have a wide variety of possible solutions, i.e. there are very many vector functions of position and time which satisfy them. Some solutions are rather simple and we will for the most part discuss only those in this chapter.

If one changing charge or current distribution gives rise to fields which satisfy the equations, and a second distribution produces different fields (which must also satisfy them), the resultant fields of the two distributions are the sums of the separate E and B fields, and the sums in turn obey the wave equations. The principle of superposition, first discussed as Equation (1.6), holds for time-varying electromagnetic fields in the same way as it does for static electric and magnetic fields.

In order to investigate fields described by Equation (11.6), let us use Cartesian coordinates and write the vector E in terms of its components as E = exEx + eyEy + ezEz. Equation (11.6) then separates into three equations obeyed by the components separately. The equation obeyed by the component Ex is



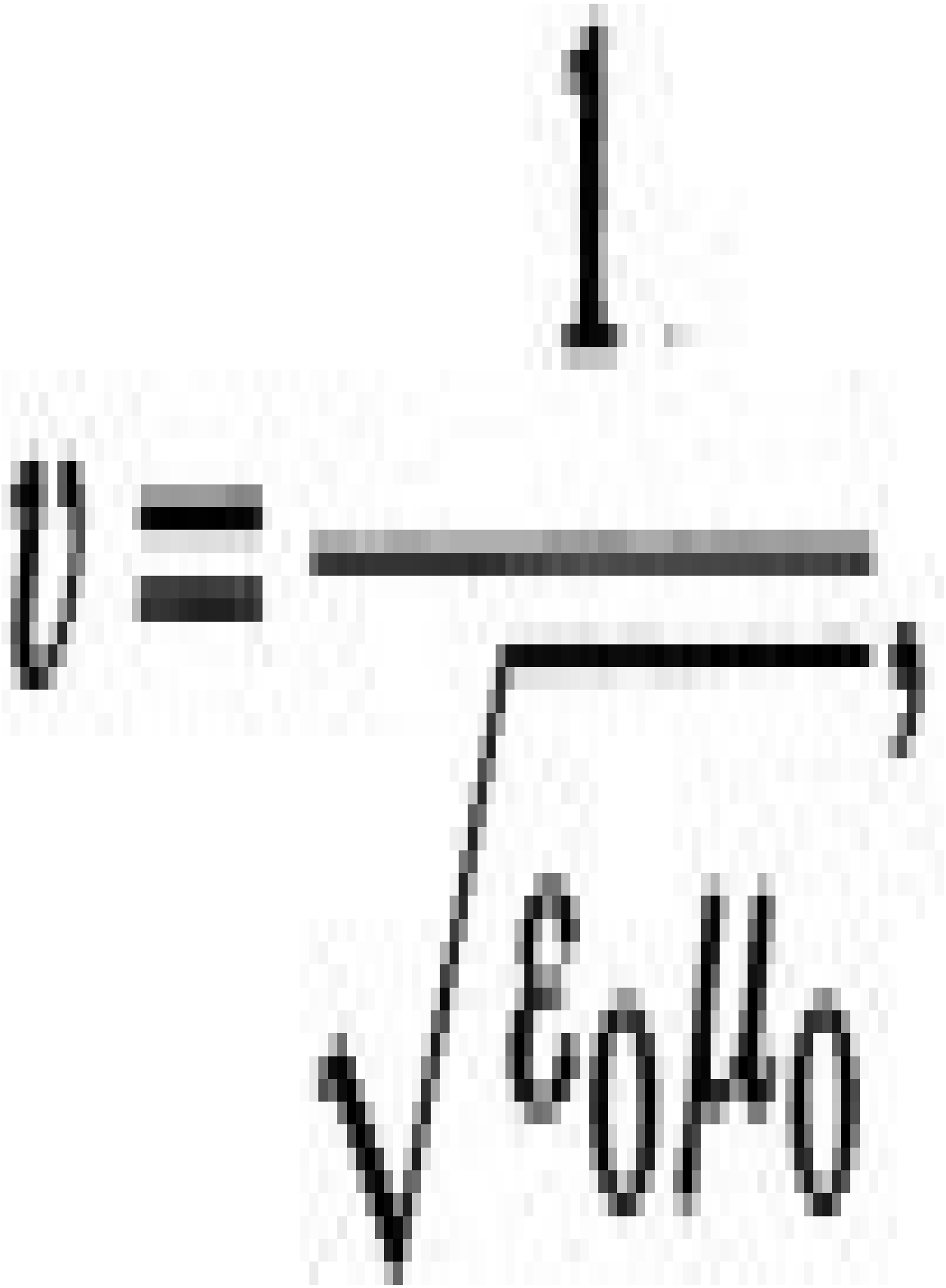
Consider the simple situation when the field component Ex does not vary in the x- or y-directions. The above equation then reduces to



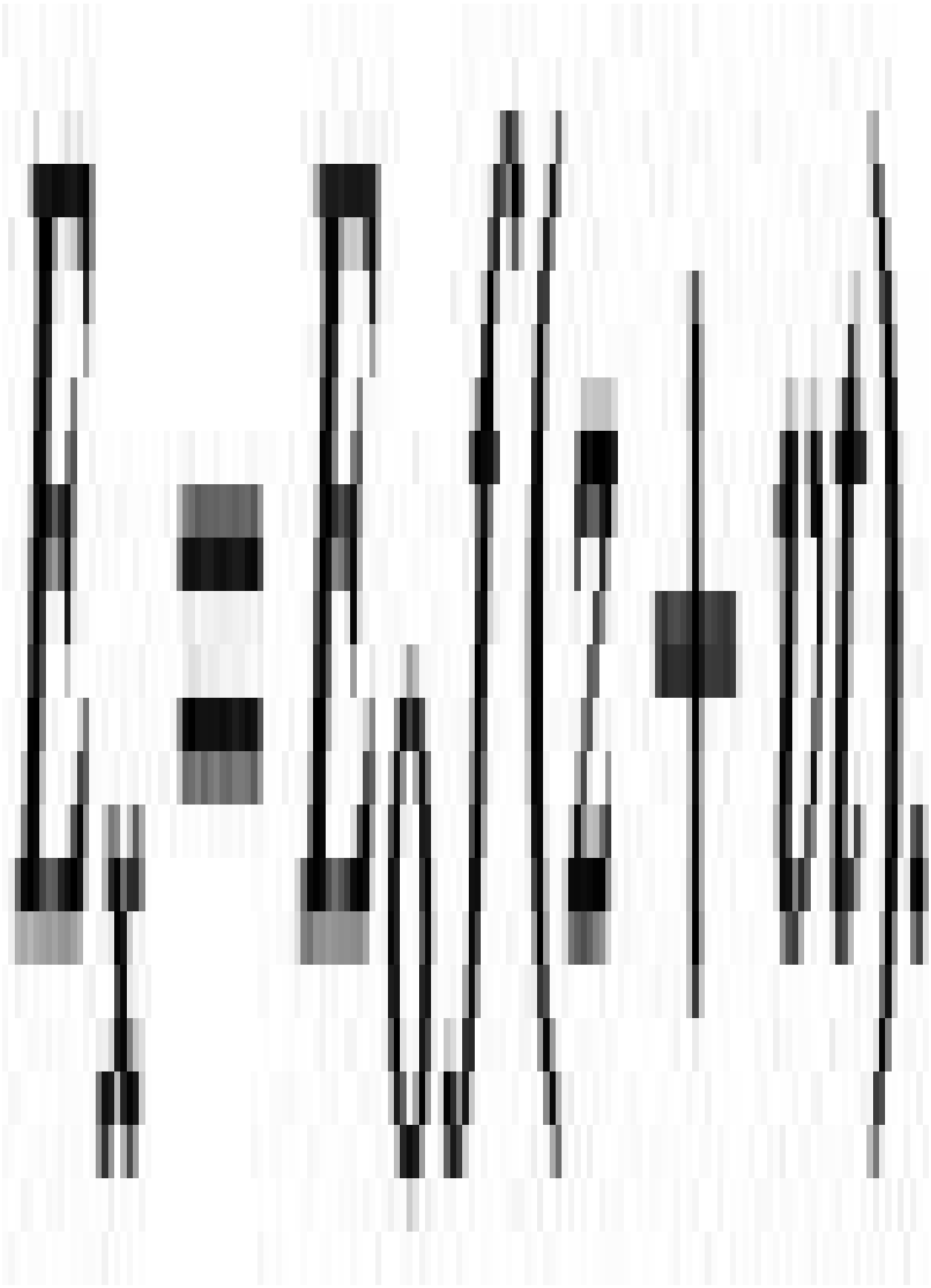
One solution to this equation has the form

(11.7)

where E0 is a constant, f is any function, and υ is given by

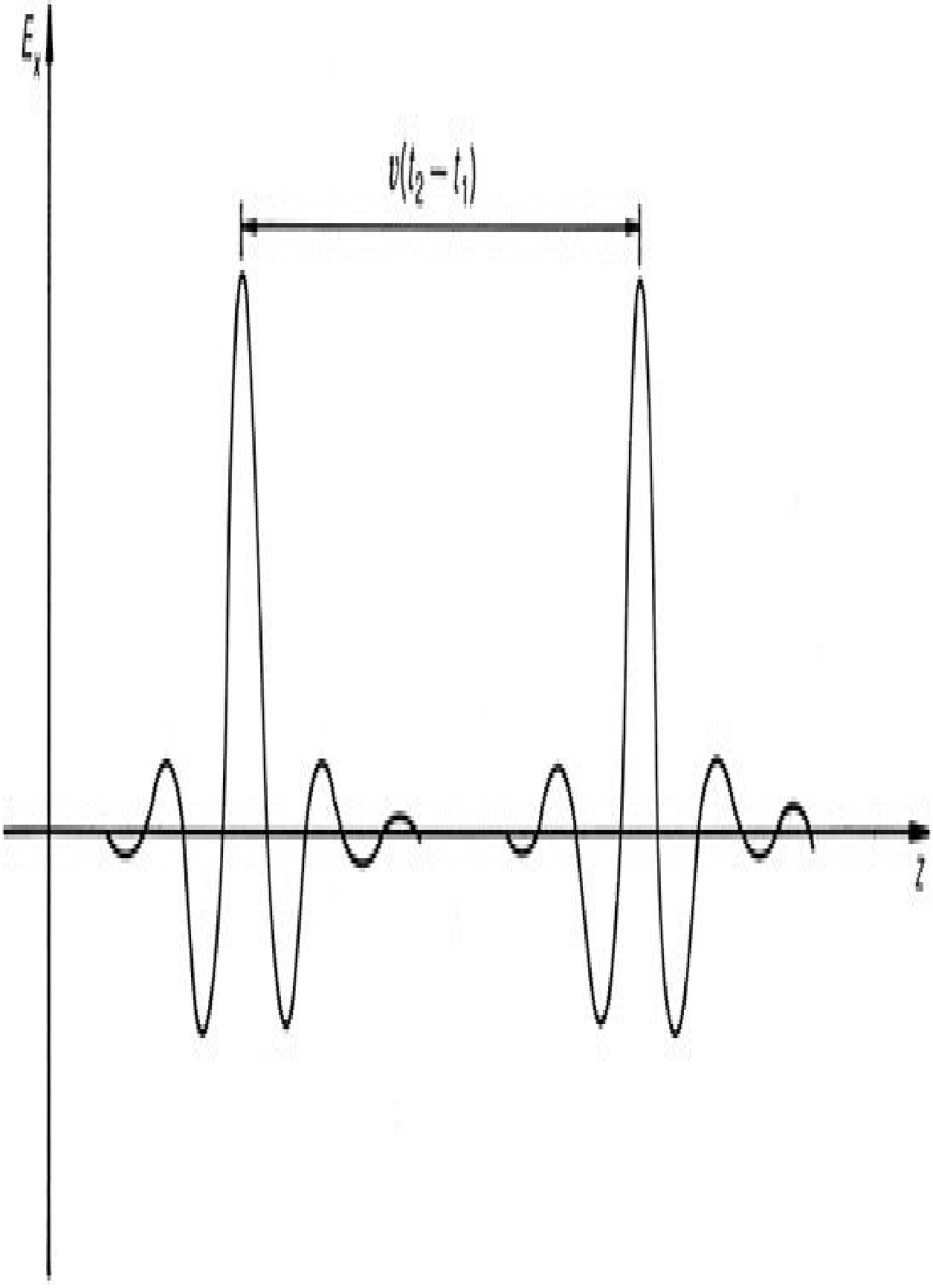


as can readily by verified by substitution. The expression (11.7) describes how the field component Ex varies with distance z and time. Suppose that at time t1 the field Ex is as shown in Figure 11.1, i.e. the function f corresponds to the profile or waveform shown. At a later time t2 it can be seen from expression (11.7) that the profile has moved along the z-axis, unchanged in shape, by an amount v(t2 – t1). Hence Equation (11.7) represents a waveform moving unchanged in shape with speed v along the positive z-axis. A similar waveform moving in the direction of the negative z-axis is given by



The three-dimensional Equation (11.6) also represents waves travelling with speed equal to $(\varepsilon_0\mu_0)^{-1/2}$. All electromagnetic waves in free space move with the same speed. This speed is given the special symbol c, and

Figure 11.1. A waveform moving along the z-axis, unchanged in shape, with speed υ. The component Ex of the electric field in the wave is plotted against coordinate z. Ex is independent of the coordinates x and y.


E_x
$v(t_2 - t_1)$
z

(11.8)

The speed c is a fundamental constant which, in agreement with the defined values of the permittivity and permeability of free space, is defined to have the value 299792458 m s−1. Prior to 1983, when the new standards of length and time were introduced, the speed c was measured in terms of the standard metre and the standard second defined separately. It was measured experimentally in a variety of ways with radiation of widely different frequency. The frequency range extended from the lowest radio frequencies (~100 kHz), through the region of frequencies corresponding to visible light, and up to x-ray and γ-ray frequencies in excess of 10^{18} Hz. All the measurements were consistent with a constant value for c, showing that all of these radiations are electromagnetic waves. The demonstration that light was electromagnetic radiation was one of the first triumphs of Maxwell's formulation of the field equations. The acceptance of the constancy of c as a natural law underpins the new system of units, which is discussed in Appendix A.

The constant c has a very special place in present physical theories. A basic postulate of the theory of special relativity is that all observers in uniform relative motion in free space measure the speed of light to have the same value c. This suggests that Maxwell's equations are already in a form obeyed by all observers in uniform motion in free space. The people in the different laboratories moving uniformly with respect to each other, if observing the same phenomena, would observe different electric and magnetic fields, but Equations (11.1) to (11.6) would still be obeyed by the fields. Maxwell's equations and relativity are discussed in Chapter 14.

11.2 PLANE WAVES AND POLARIZATION

The simplest type of wave that is a solution to Equation (11.5) and (11.6) is a plane wave. Plane waves will be our main concern in this chapter. Far enough away from any source of radiation, the electromagnetic wave is a plane wave, to a good approximation, and the study of plane waves illustrates many important features of the electromagnetic field. We will first define plane waves and then discuss what is meant by the polarization of an electromagnetic wave.

Suppose that an electromagnetic field is produced by a current in a coil, and that the current oscillates continuously with constant amplitude and constant angular frequency ω. In the steady state the field strengths everywhere will oscillate at the same angular frequency. The electromagnetic waves produced are said to be monochromatic. (The name is taken from the visible part of the spectrum, but it is usually applied to waves that oscillate sinusoidally at any fixed frequency.) The power radiated by the coil, averaged over a complete cycle of the current, is constant in the steady state. The power is carried away by the radiated waves, and the same average power crosses any sphere whose centre is at the position of the coil. If we consider two such spheres, one closer in than the other, the energy crossing unit area per second (in a direction along a particular line drawn away from the coil) over the inner sphere is greater than the energy crossing unit area per second in the same direction over the outer sphere. The average energies depend on the amplitudes of the electric and magnetic fields at the appropriate points on the two spheres, and hence the field strengths decrease with increasing distance from the source.

The fields at two points are said to be in phase if their oscillations occur in unison. A wave front is a surface over which the fields are in phase. Wave fronts in which the fields have a constant phase spread out from the coil and become weaker, as the fields decrease, like the ripples spreading out from a disturbance in the middle of a pond. Far away from the coil, in a region whose dimensions are small compared with the distance from the coil, the wave fronts become very

nearly plane surfaces. These planes are perpendicular to the direction of propagation, which is the line joining the region of interest to the far-off coil. The diminution in the amplitudes of the fields within the region of interest is negligible, and the wave is said to be a plane wave. A plane wave has its wave fronts parallel to each other, and the electric field in a plane monochromatic electromagnetic wave propagating along the positive direction of the z-axis, with the electric field lying in the direction parallel to the x-axis, can be represented by the real part of the expression

(11.9)

The field has the same amplitude E0 everywhere, and at all points with the same z-coordinate, i.e. all points in any plane parallel to the x–y plane, the field has the same phase. It is oscillating with angular frequency ω. It can readily be shown that the expression (11.9) is a solution to the wave Equation (11.6), and hence, repeating the arguments which followed Equation (11.7) in the last section, the speed of propagation of the wave, c, is given by

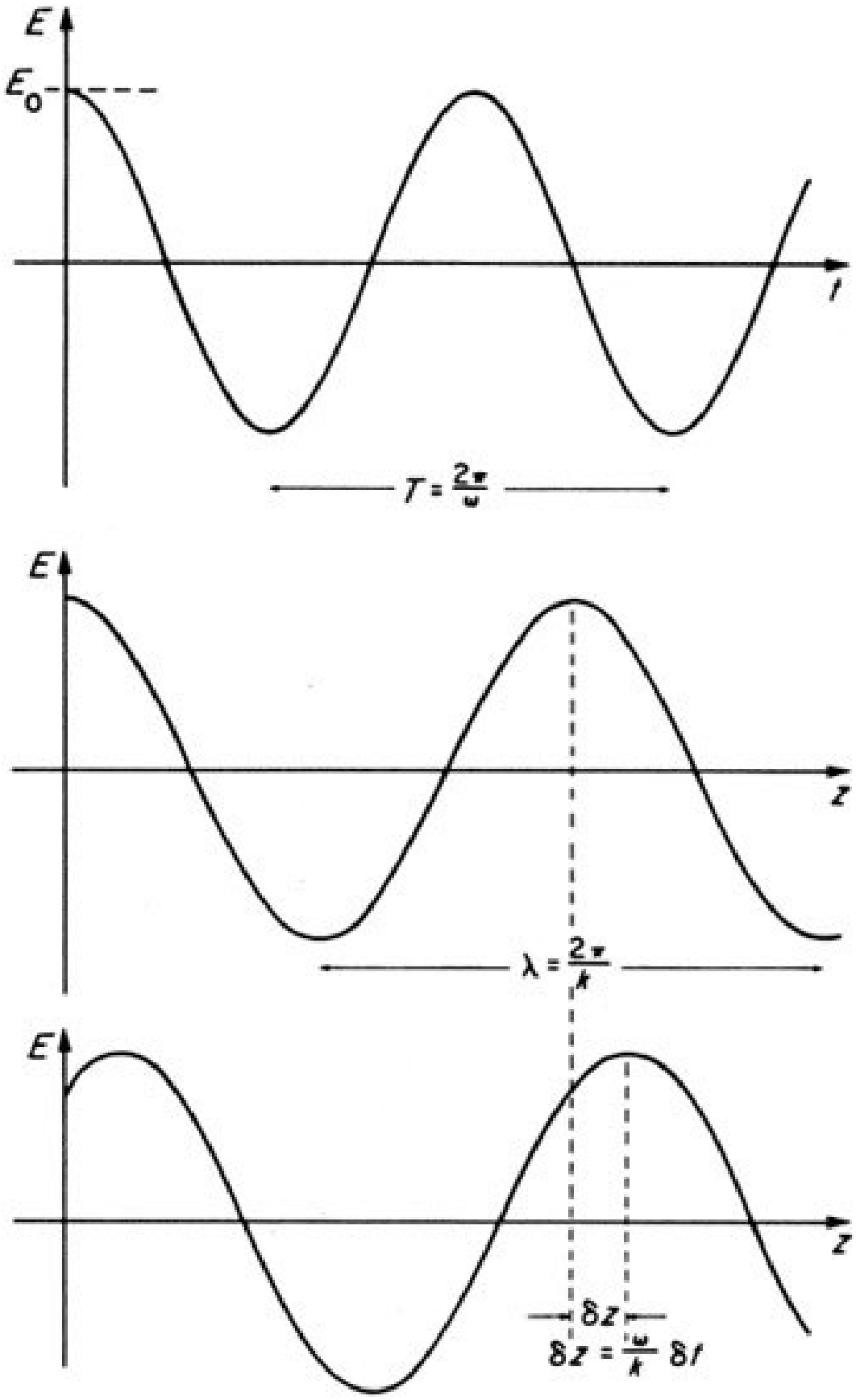
(11.10)

k is called the wavenumber, and is measured in m−1. Figure 11.2 shows how the field varies with time at a fixed position z = 0, and how it varies with position at time t = 0. The variation with distance is sinusoidal, and the distance between equivalent points on successive cycles is equal to the wavelength λ of the wave,

(11.11)

Figure 11.2. The top figure shows how the electric field in the plane wave (11.9)

varies with time at all points with coordinate z equal to zero. The field is independent of coordinates x and y. The centre figure shows how the field varies with z at time t = 0. The bottom figure shows the variation with z at a later time δt. The wave crests have moved forward a distance δz equal to $\omega \, \delta t/k$.


E
E_0
t
$T = \frac{2\pi}{\omega}$
E
z
$\lambda = \frac{2\pi}{k}$
E
z
δz
$\delta z = \frac{\omega}{k}\,\delta t$

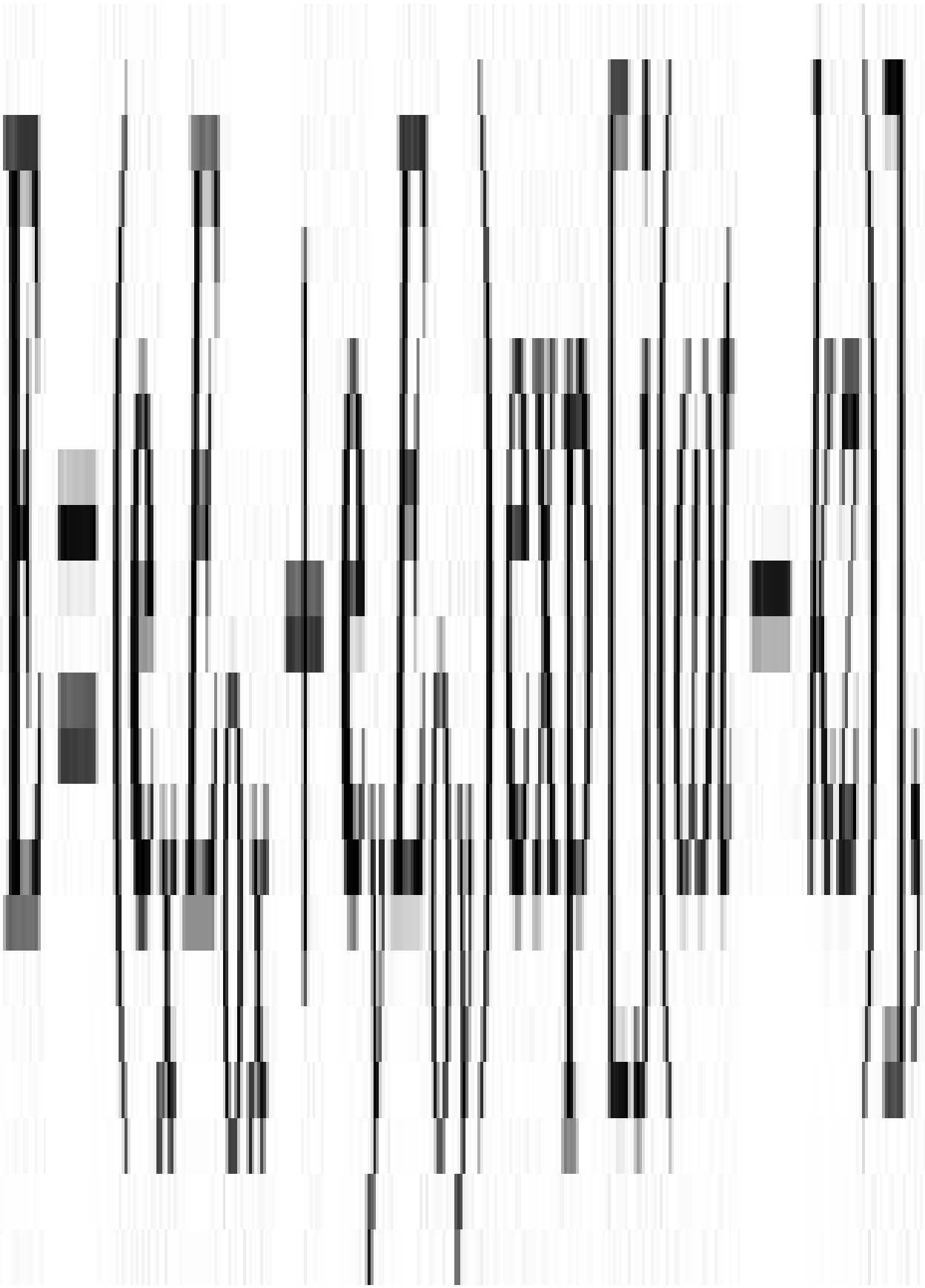
The third diagram in Figure 11.2 shows how the field varies with z at time δt. The waveform has the same shape, but the maxima have moved forward a distance δz, such that

Combining Equations (11.10) and (11.11) we obtain a relation between the wavelength λ and the frequency $\nu(= \omega/2\pi)$ of electromagnetic waves in free space,

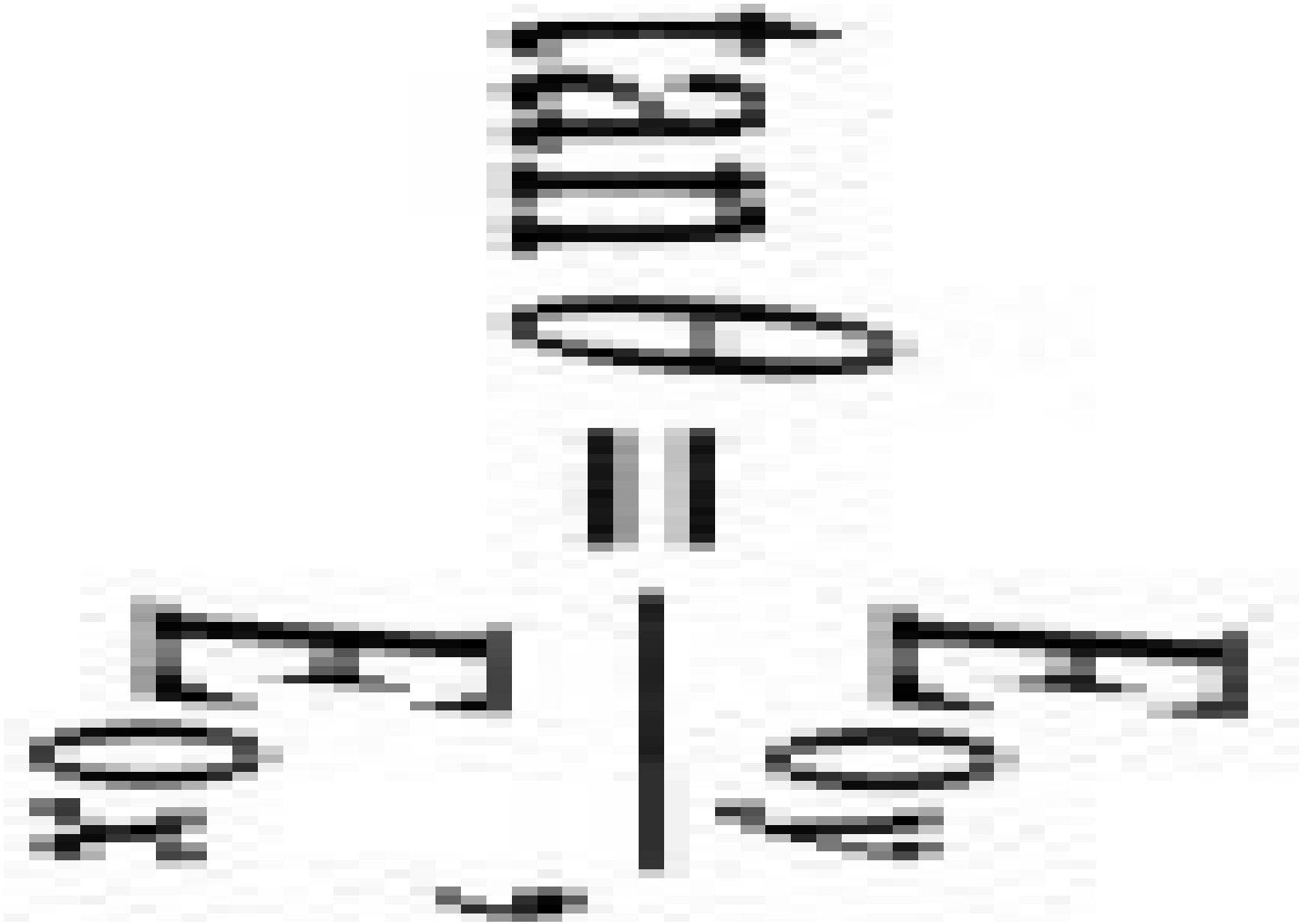
(11.12)

We will now discuss what is meant by the polarization of a plane wave. Expression (11.9) represents a linearly polarized plane wave. The electric field vector always lies parallel to the x-axis, pointing either in the positive or negative direction depending on the instant at which we observe the oscillating field. A linearly polarized monochromatic plane wave in which the electric field is always parallel to some other line in the x–y plane can be expressed as

(11.13)



the sum of two separate plane wave solutions to Equation (11.6). This corresponds to two oscillations which are in phase and polarized at 90° to each other. The resultant is a vector which oscillates along a line making an angle θ with the x-axis, where



as shown on Figure 11.3.

The plane wave

(11.14)

Figure 11.3. The electric field in a linearly polarized plane wave in which the electric field vector oscillates along a line making an angle θ with the x-axis.

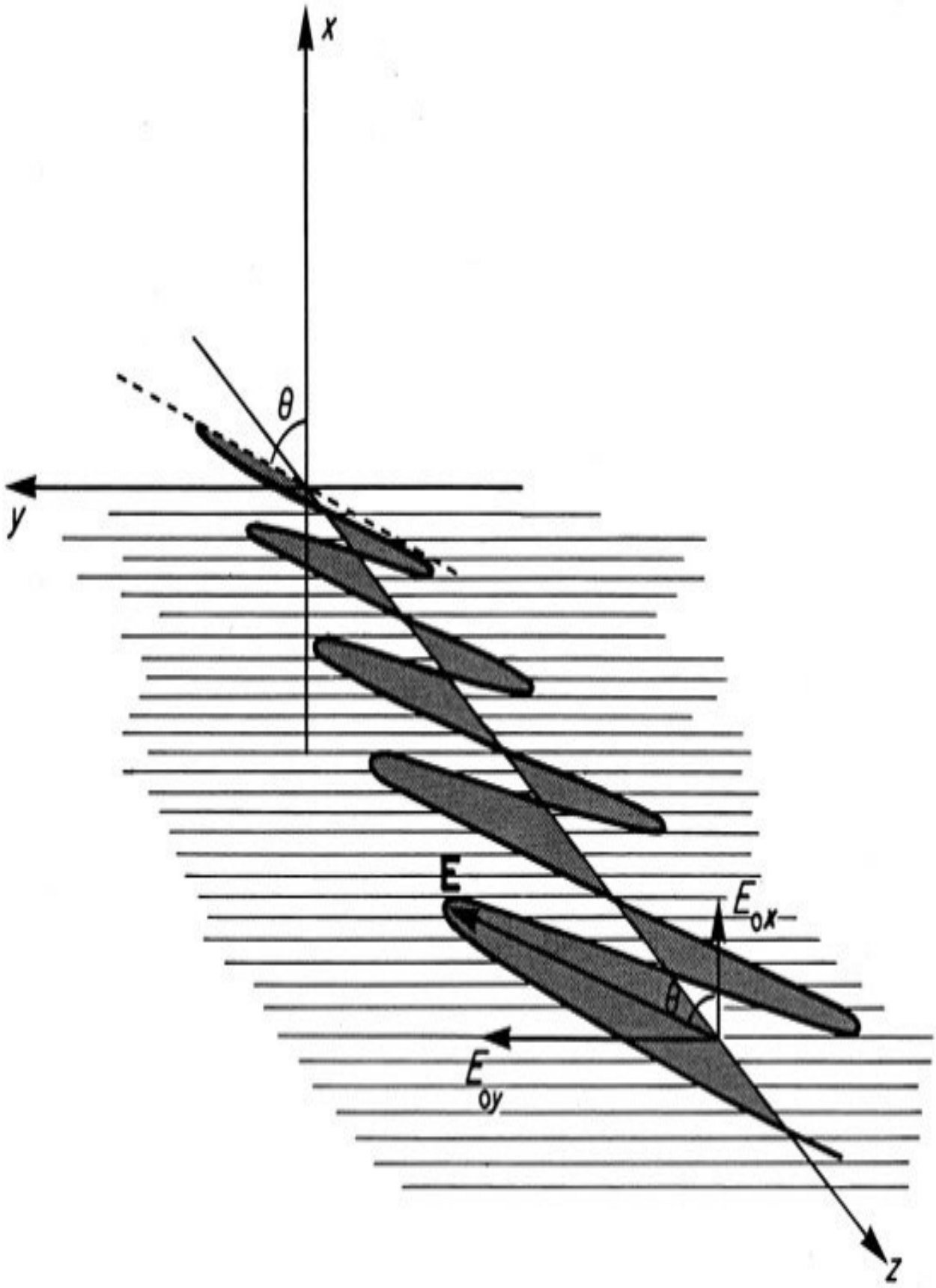
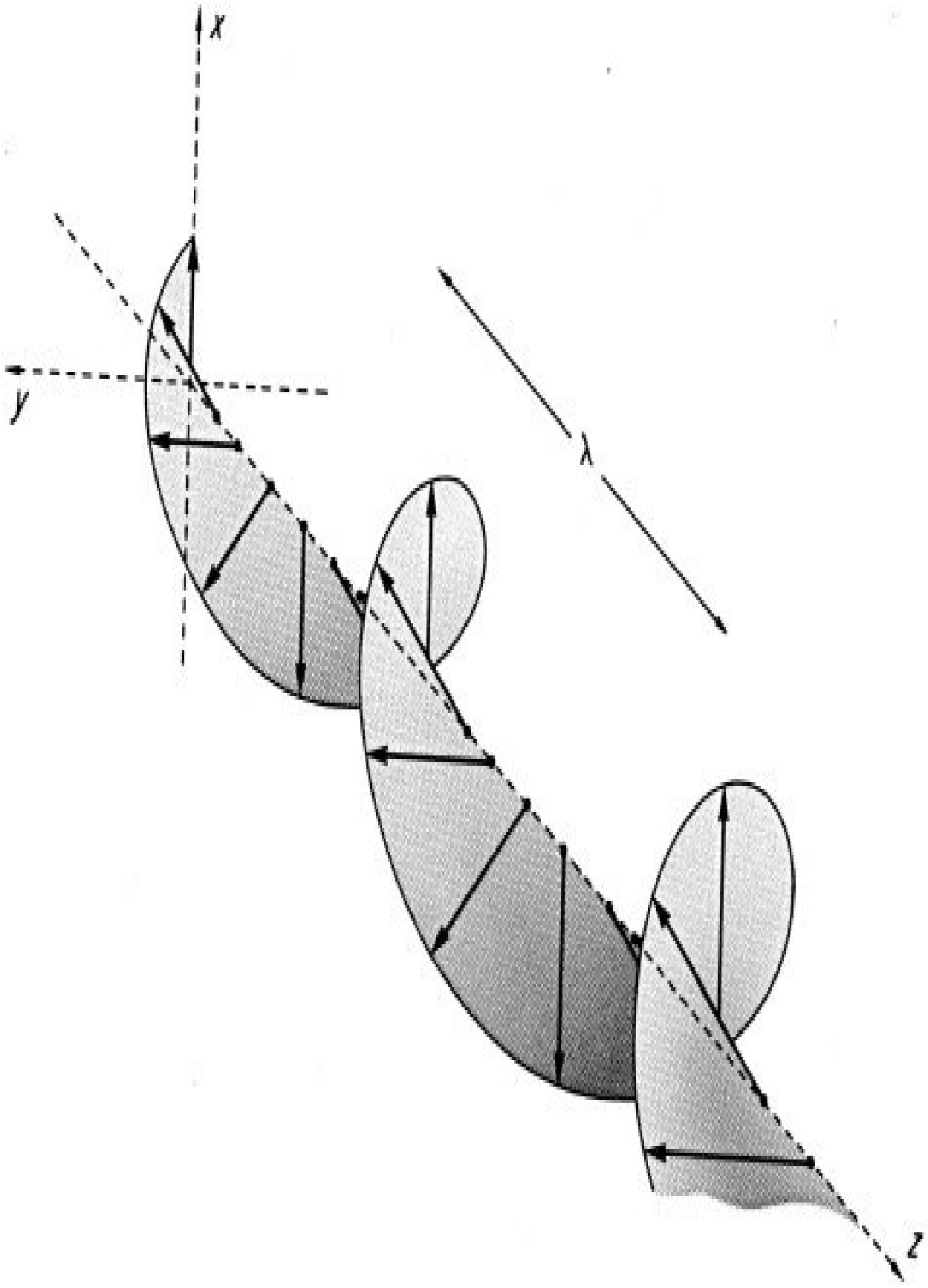
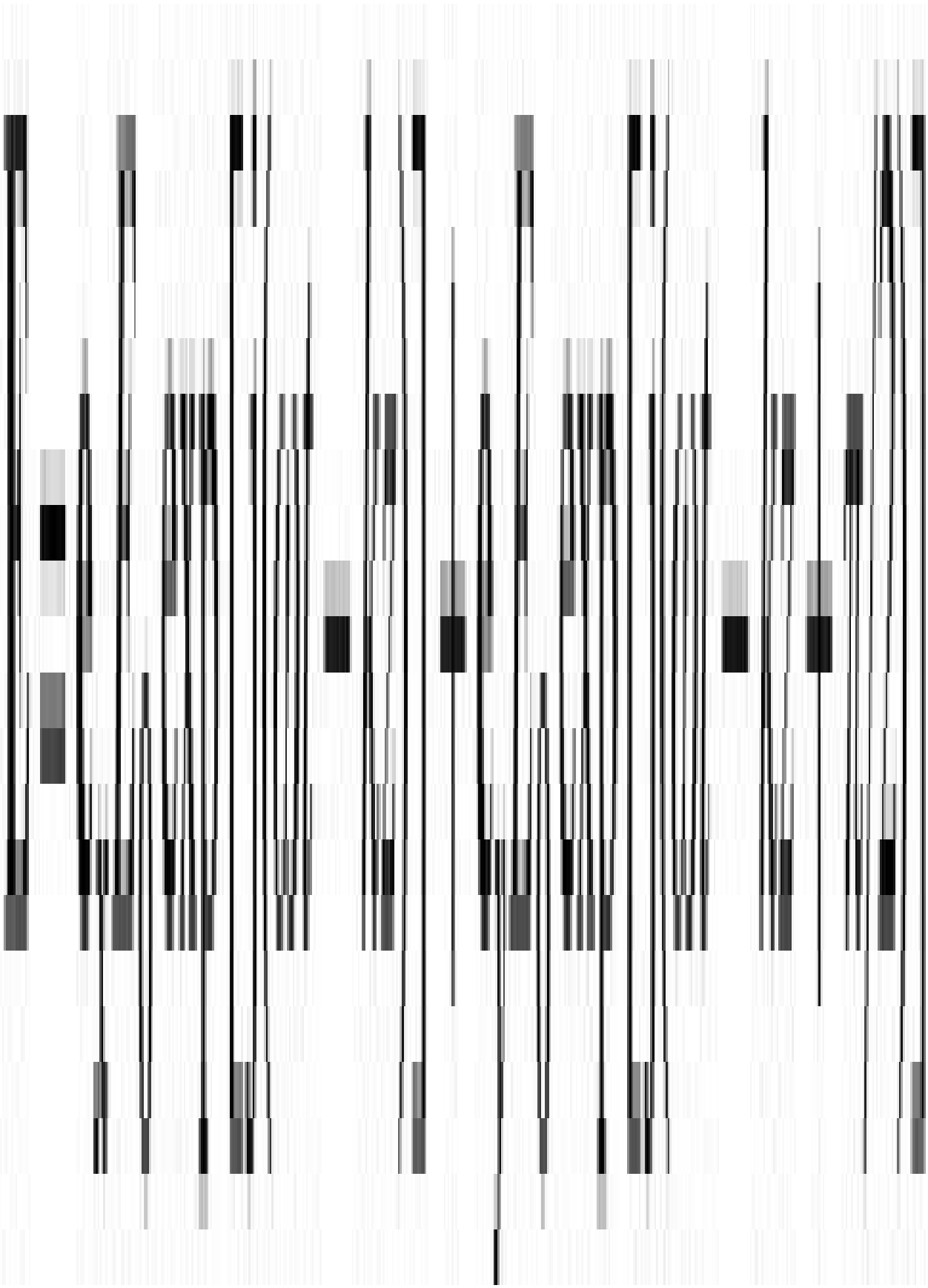

x
θ
y
E
E_{ox}
θ
E_{oy}
z

Figure 11.4. The electric field in the circularly polarized wave (11.15). The field is shown at time t = 0, and is independent of the coordinates x and y. Compare this wave with the plane polarized wave shown in Figure 11.5.


x
y
λ
z

where ϕ is an arbitrary constant phase angle, is also a solution to the wave Equation (11.6). The sum of waves given by expressions (11.14), (11.13) and (11.9) is also a plane wave solution to the wave equation. By adding monochromatic plane waves linearly polarized in different directions and out of phase we can construct expressions which represent plane waves in more complicated states of polarization. For example the field

(11.15)



represents a circularly polarized plane wave. It is a wave progressing along the z-axis in which the vector E has a constant length E0 but is continuously changing its direction. At any fixed point the vector E rotates* in a circle with angular frequency ω. The field as a function of position at the fixed time t = 0 is illustrated in Figure 11.4.

A randomly polarized plane wave (often called an unpolarized plane wave) is one in which the direction of oscillation of the field is changing randomly with time. A plane wave from most visible light sources (except lasers!) contains contributions from very many individual atomic or molecular de-excitations, and is randomly polarized. A plane wave at a frequency of 1 MHz which arises from the controlled sinusoidal variation of current along a wire has a fixed polarization. Only when the changing charge and current distributions producing the waves are controlled in a regular manner will such sources emit radiation of a fixed state of polarization.

11.2.1 Plane waves in free space

In this section we will first show that the electric and magnetic field vectors in a plane wave lie in directions perpendicular to the direction of propagation of the wave. We will afterwards derive a relation between the electric and magnetic fields in the wave.

Consider a monochromatic plane wave solution to the wave Equation (11.6) in which the electric field is given by

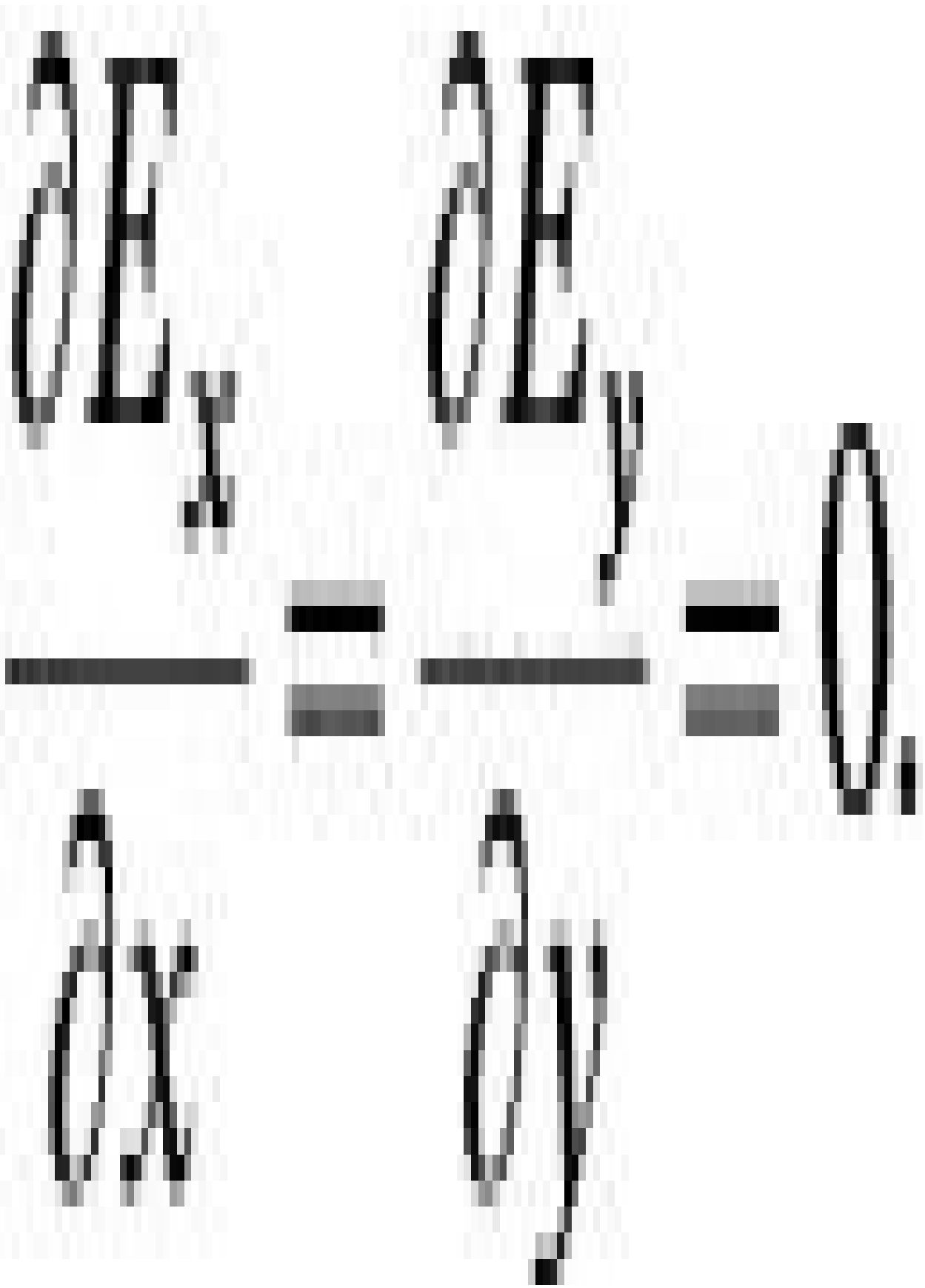
(11.16)

where $\omega/k = c$, and E_0 is an arbitrary constant vector. This expression represents a linearly polarized plane wave with the electric field direction arbitrarily chosen. We will now show that the vector E_0 must lie in a plane parallel to the x–y plane. The choice of the positive z-axis to be the direction of propagation of the wave involves no loss of generality, since we are free to choose the coordinate axes in whatever directions are most convenient.

The field given by Equation (11.16) must satisfy all Maxwell's equations. Hence from Equation (11.1)

(11.17)

at all points, and at all times. For the plane wave (11.16) the field is independent of the coordinates x and y, and hence

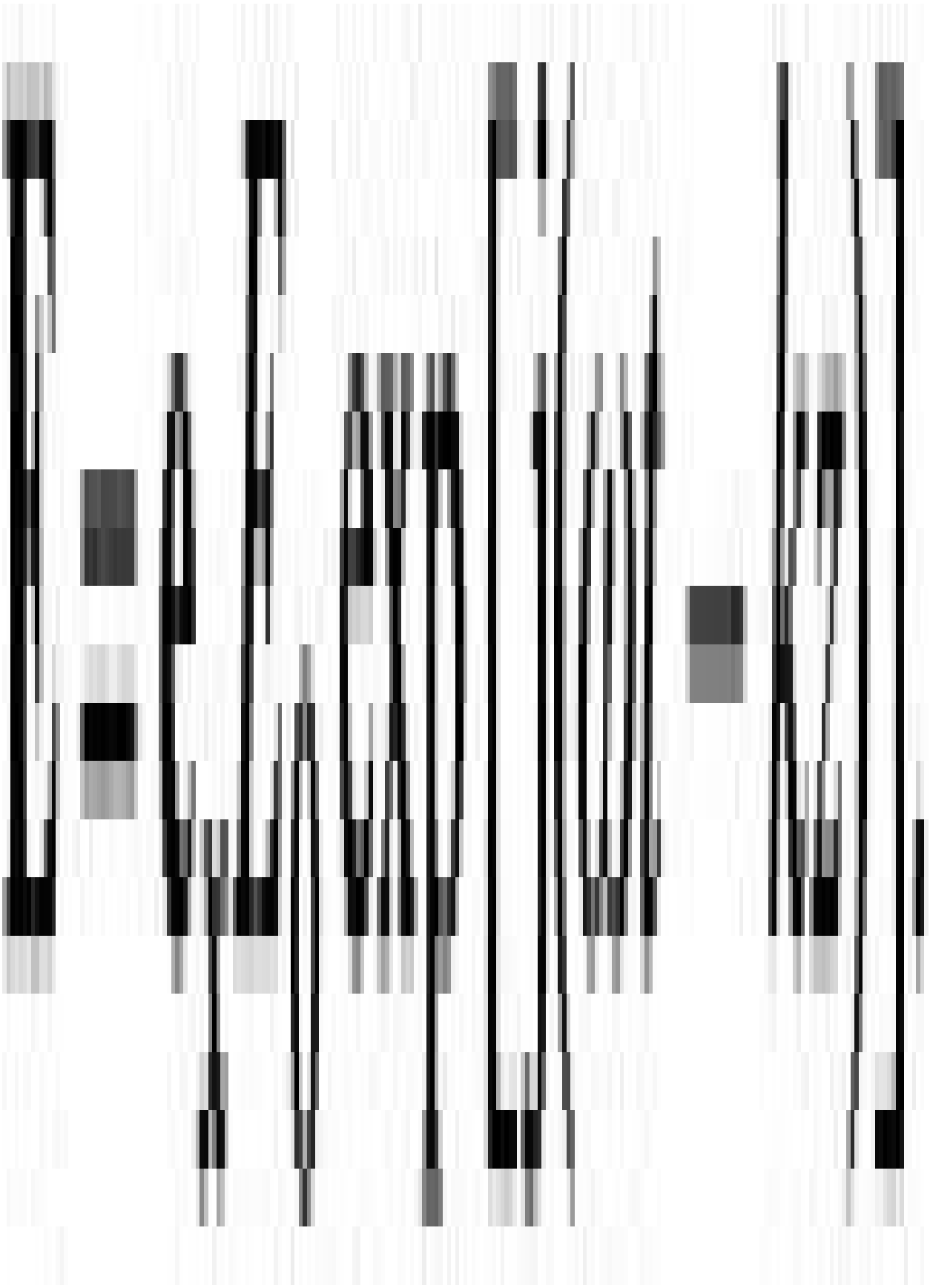


Equation (11.17) now becomes

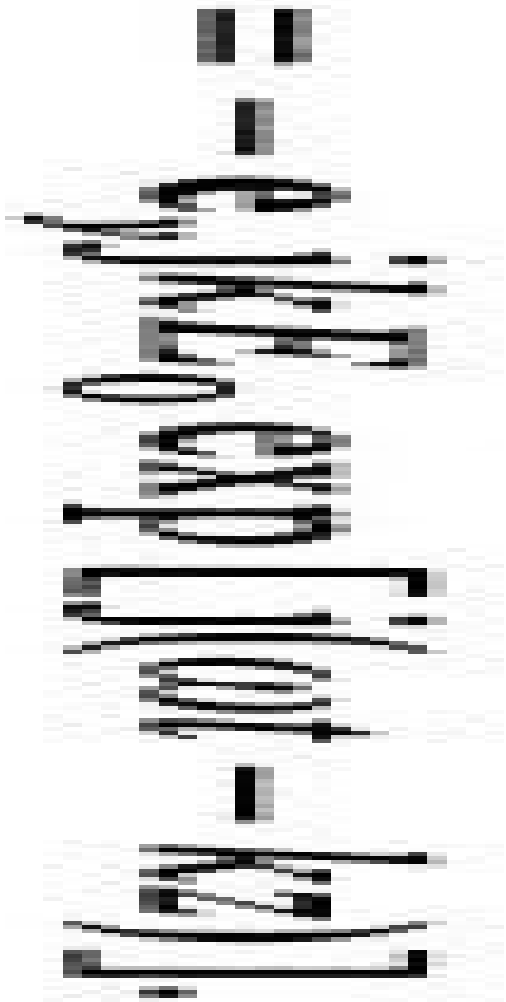
and this tells us that the component Ez is independent of the coordinate z. The only way this statement can be reconciled with the expression (11.16) for the field E, is for the component Ez to be everywhere zero. Hence the vector E0 must be perpendicular to the direction of propagation.

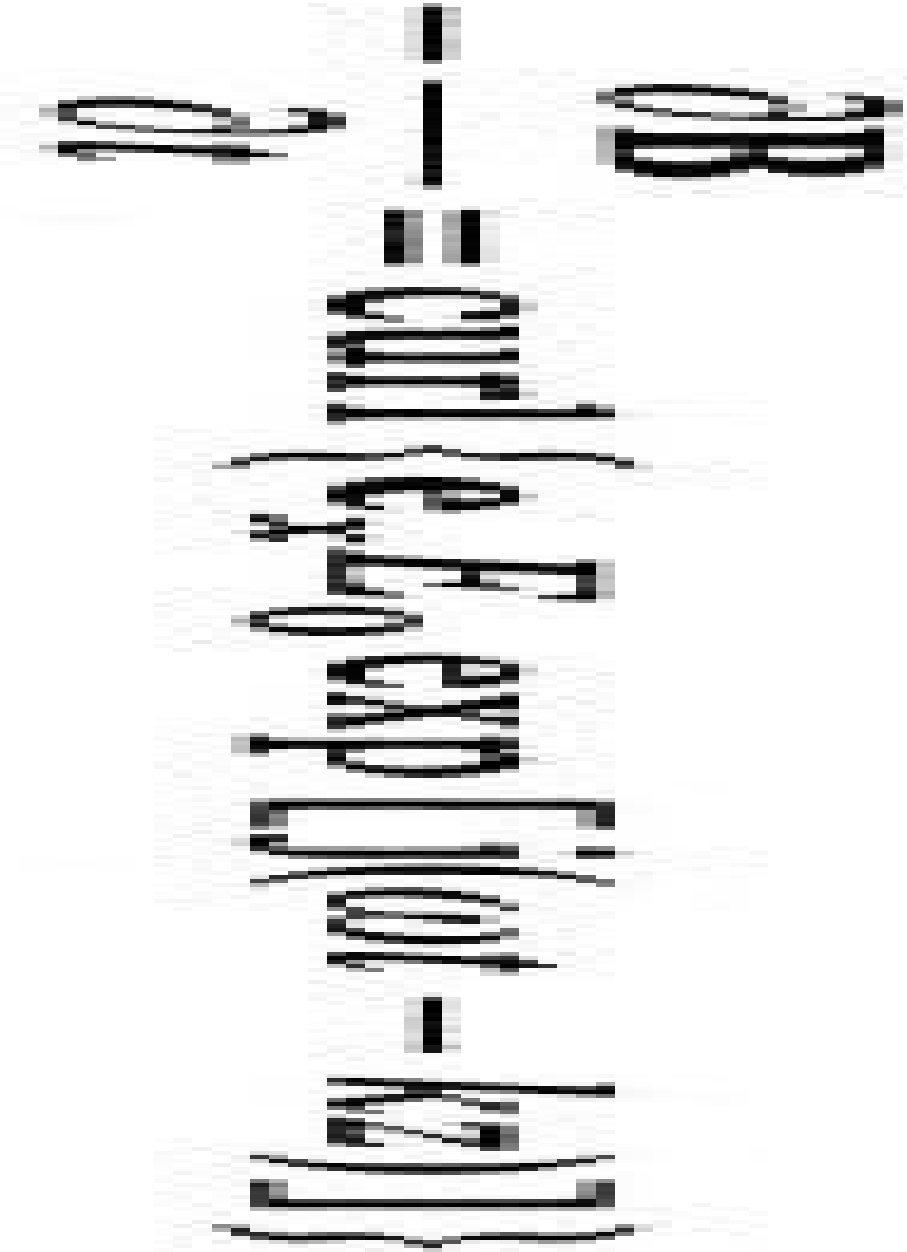
This conclusion is true for any plane wave, not only a linearly polarized plane wave. A wave in a more complex state of polarization can be expressed as the sum of linearly polarized waves with different amplitudes and phases, and the proof given above holds for each term in the sum, and hence for the whole wave. The electric field vector in a randomly polarized plane wave must also lie in directions perpendicular to the direction of propagation of the wave. The average time interval between the random changes in direction of the polarization of the wave is much greater than the period. Thus one is justified in using the expression (11.16) in the time intervals between changes, and the proof given above remains valid.

Let us now turn to the problem of determining the magnetic field B in a plane wave in terms of the electric field E. Let the electric field be given by expression (11.16), in which the vector E0 lies in the x–y plane. For simplicity we will choose the direction of the field E to be along the x-axis, i.e.

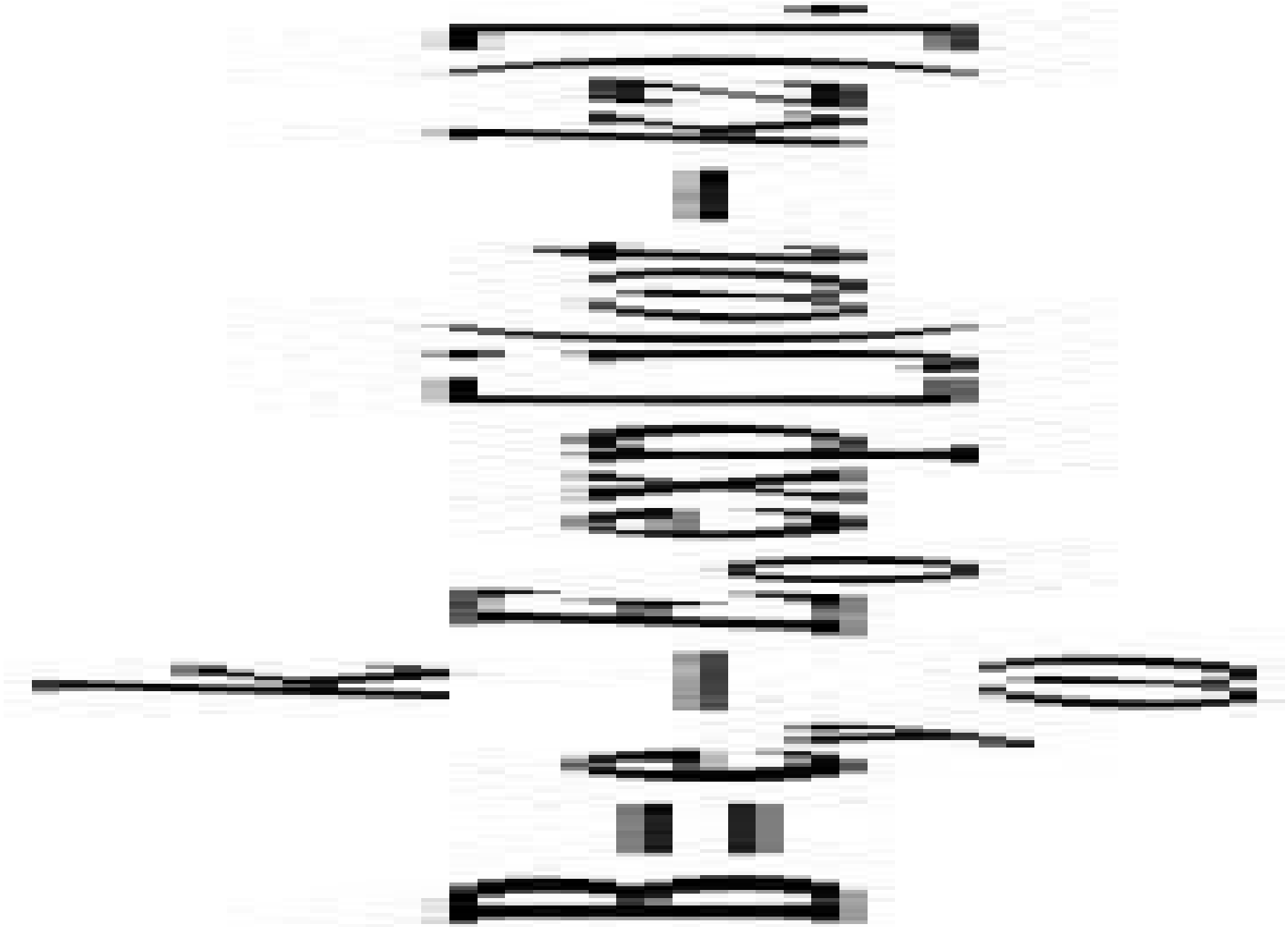


On substituting this into Equation (11.4) we obtain

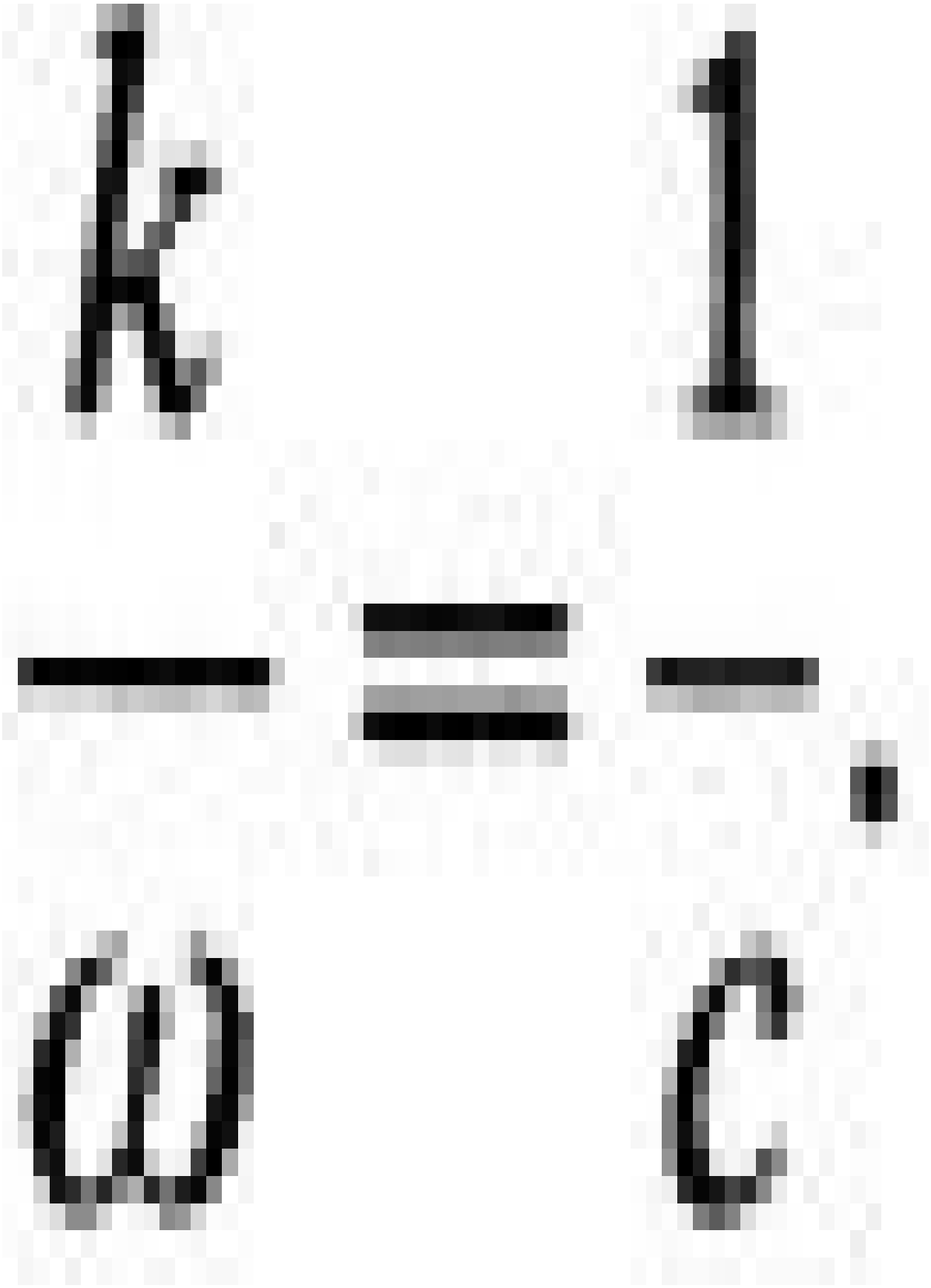




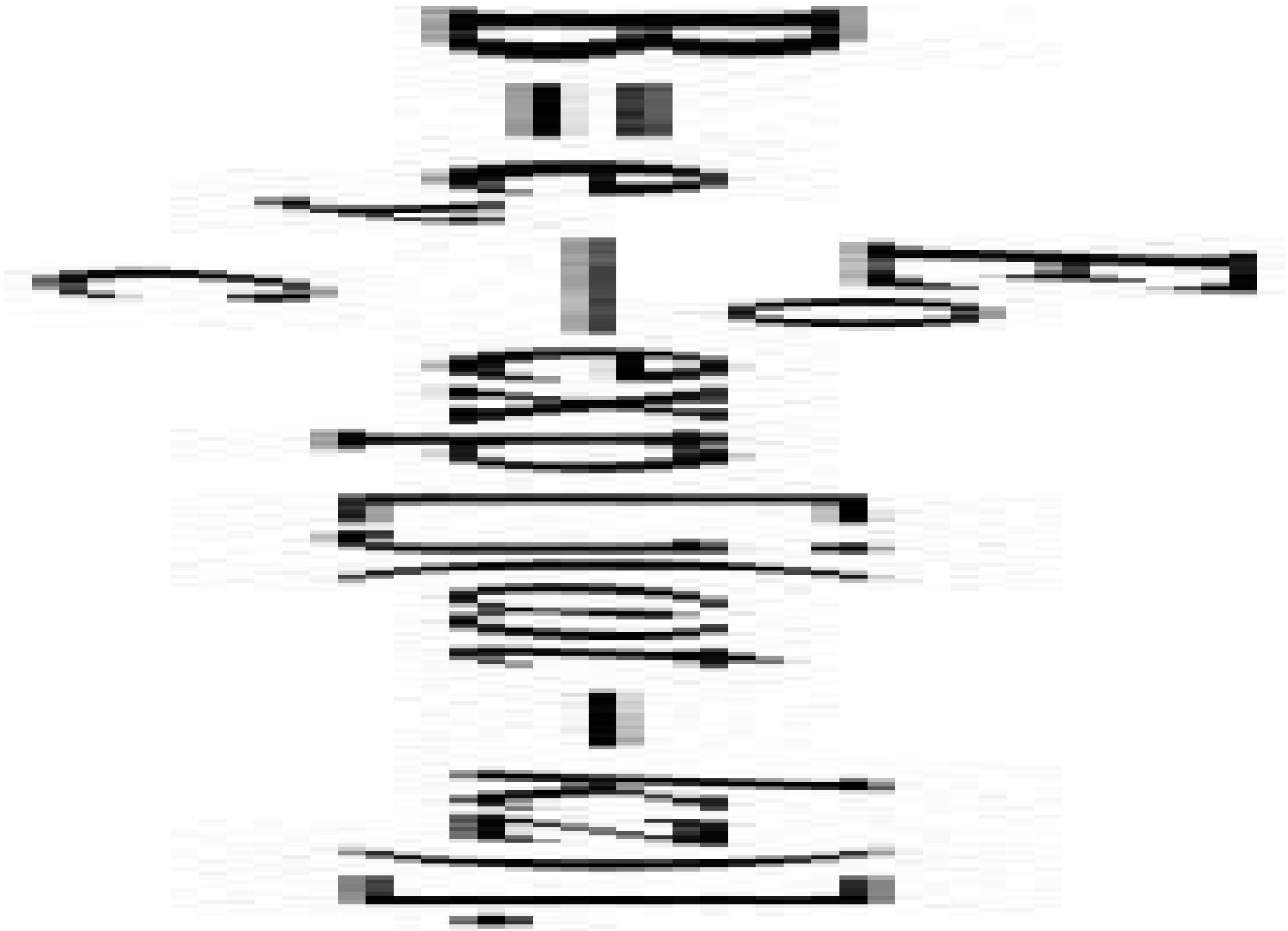
Integrating this equation with respect to time we find



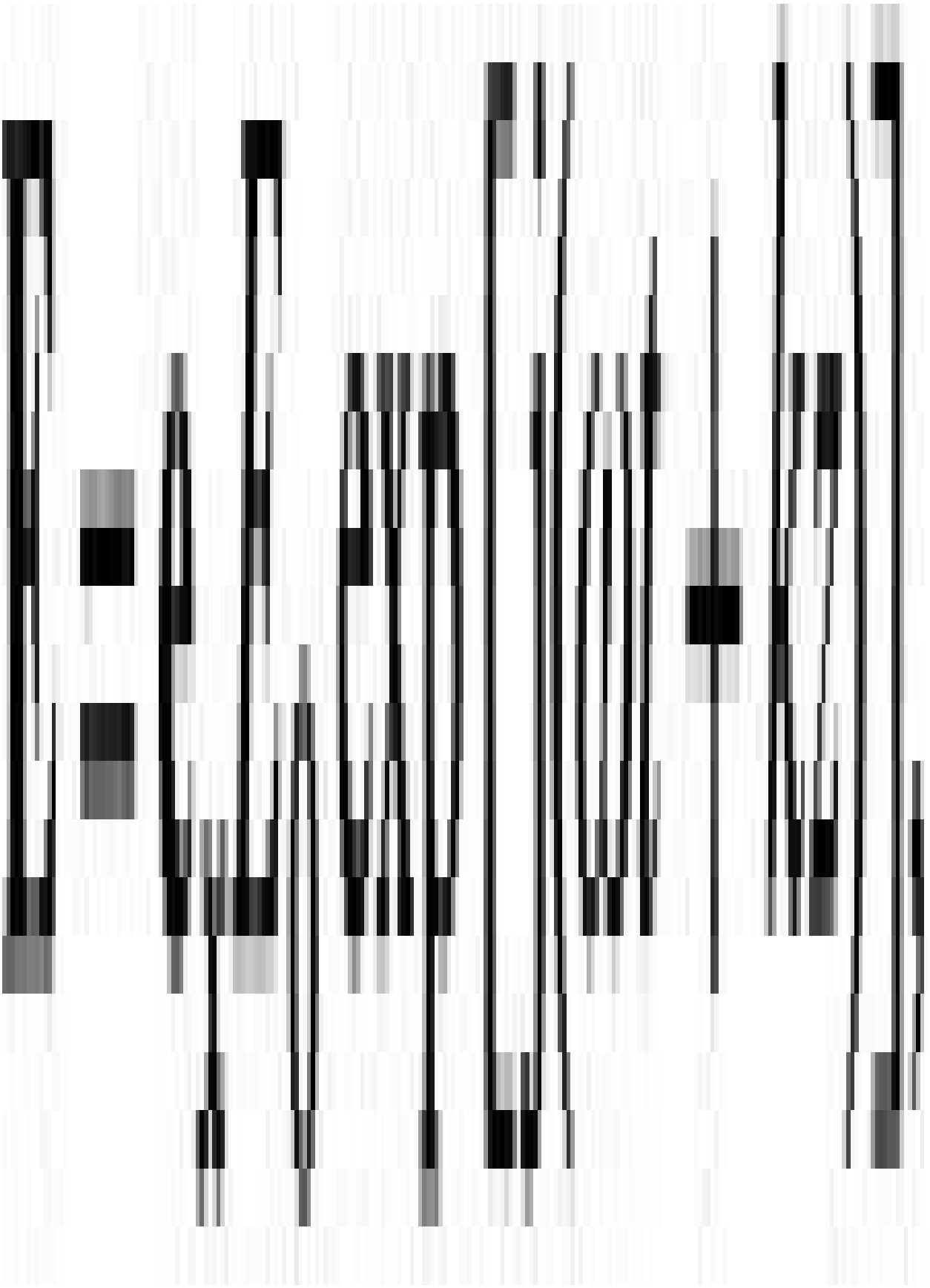
There is no constant of integration in the expression for the field B in the absence of a steady field. From Equation (11.10) the ratio of the wavenumber k of the wave to its angular frequency ω is the reciprocal of the speed of the wave



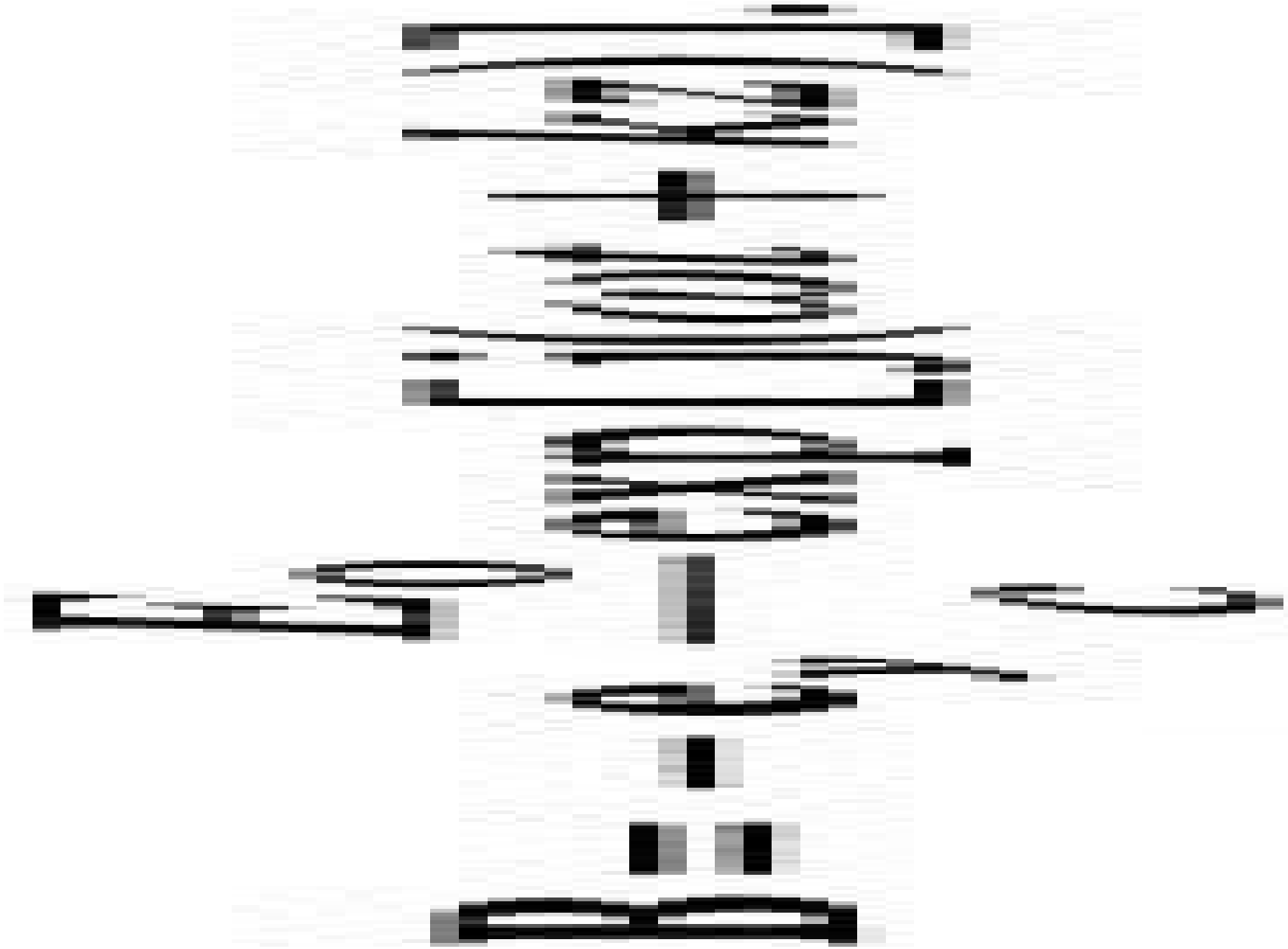
Hence, finally,



If a similar analysis is pursued with the wave progressing in the negative z-direction, the field E is given by



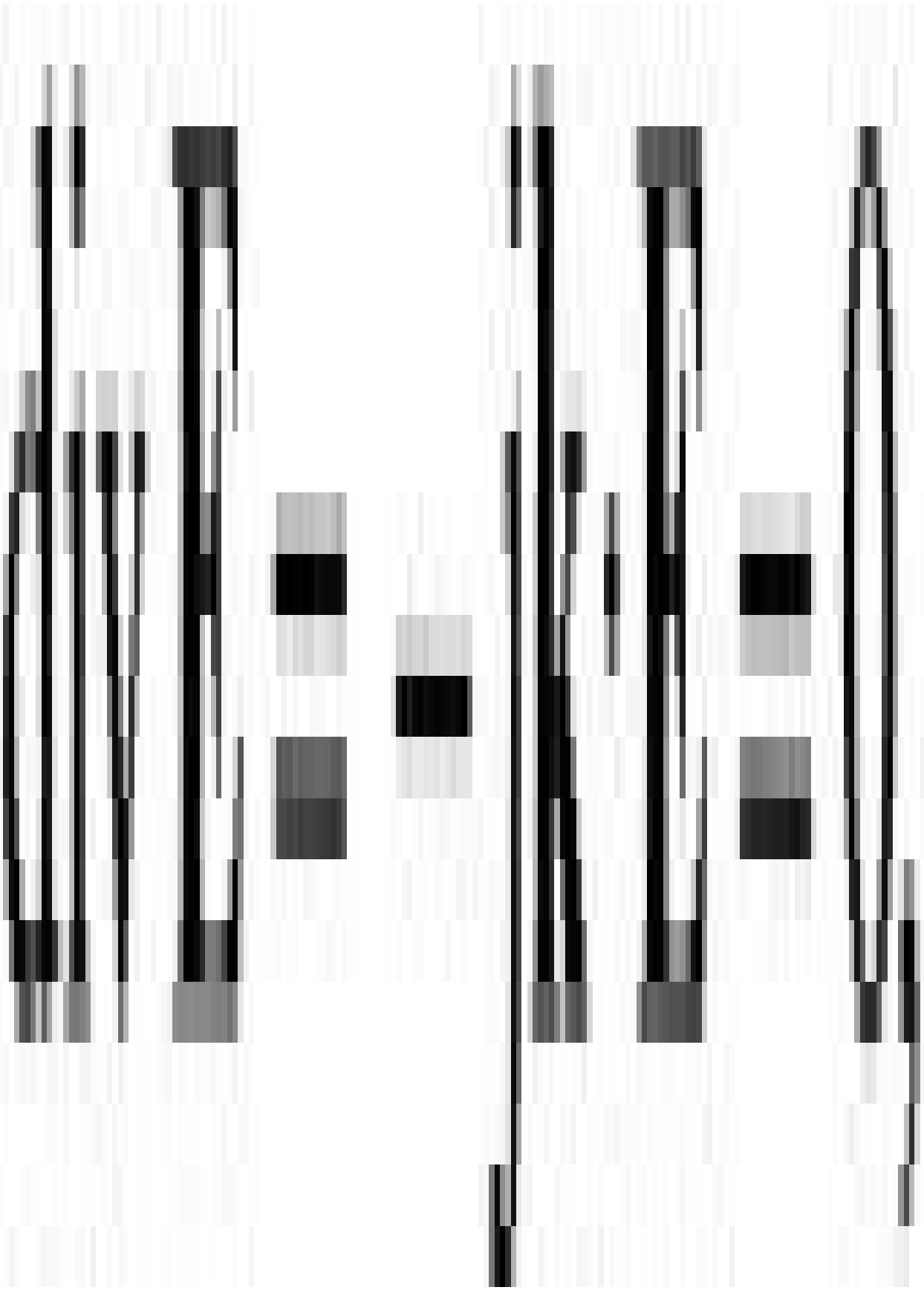
and the field B can be shown to be



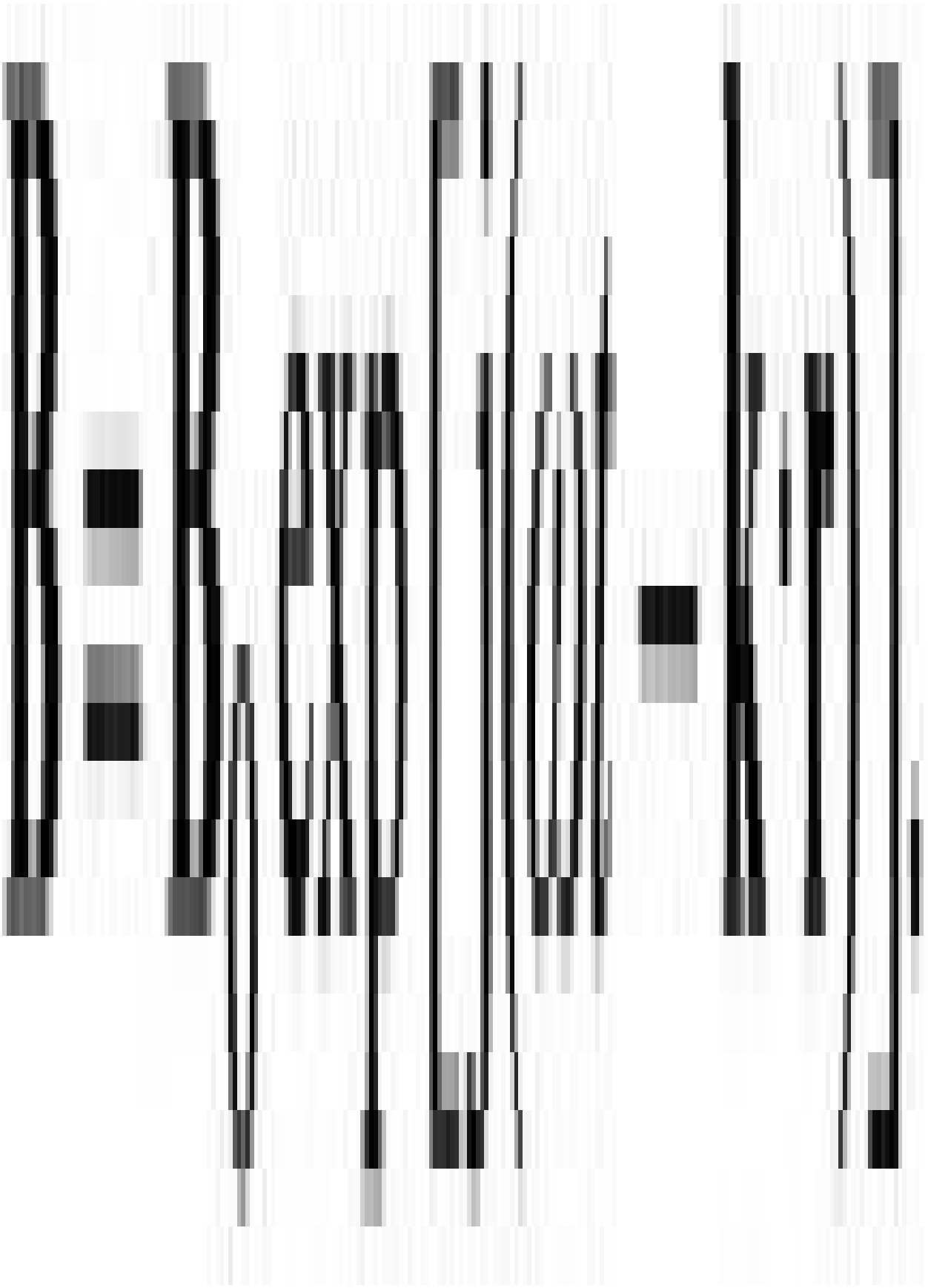
The general expression for the linearly polarized plane wave solution to Equation (11.6) is of the form

(11.18)

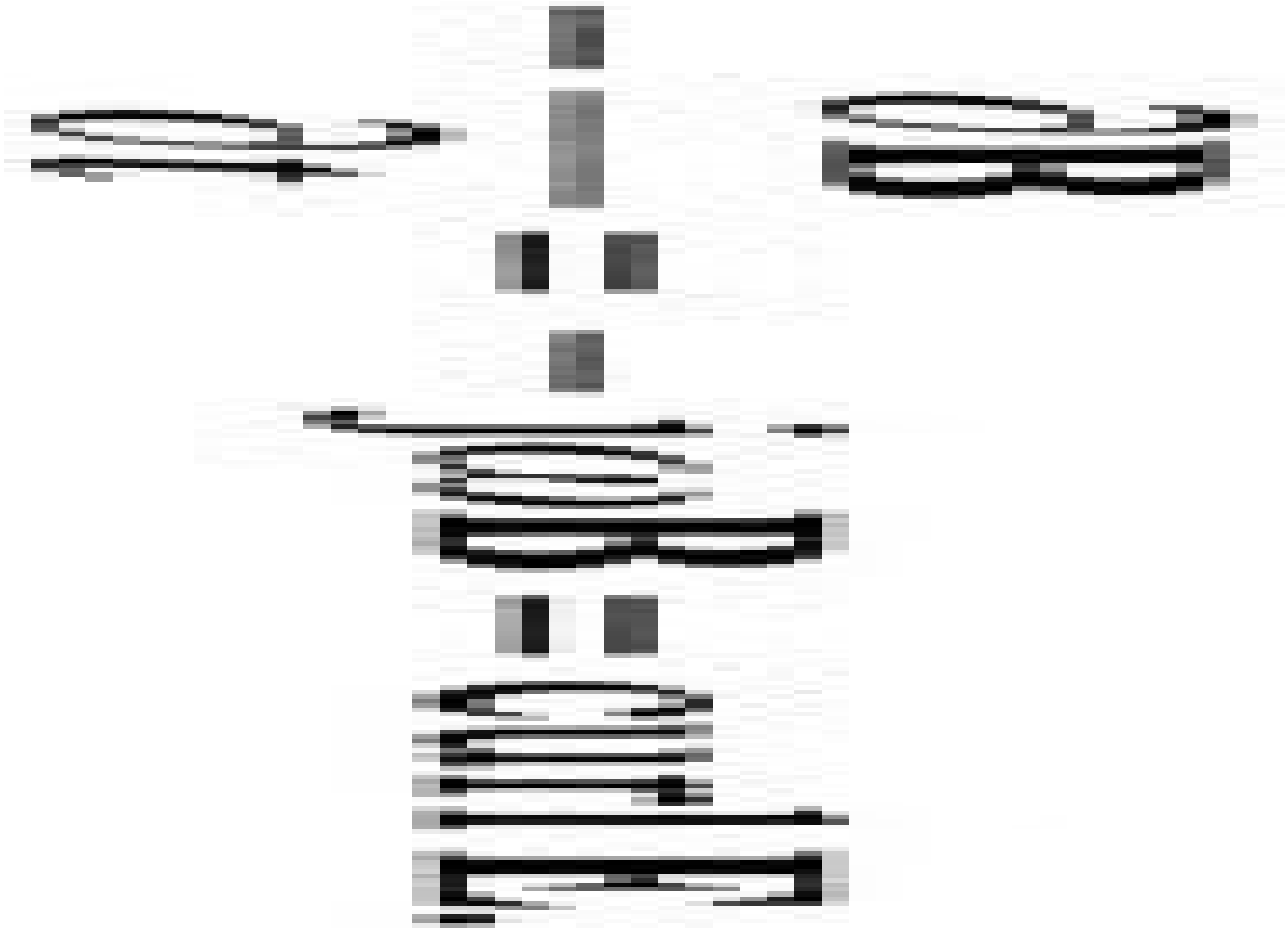
where the vector k is called the propagation vector and is in the direction of propagation of the wave with magnitude equal to the wavenumber k. Evaluating the divergence of this field and equating it to zero gives



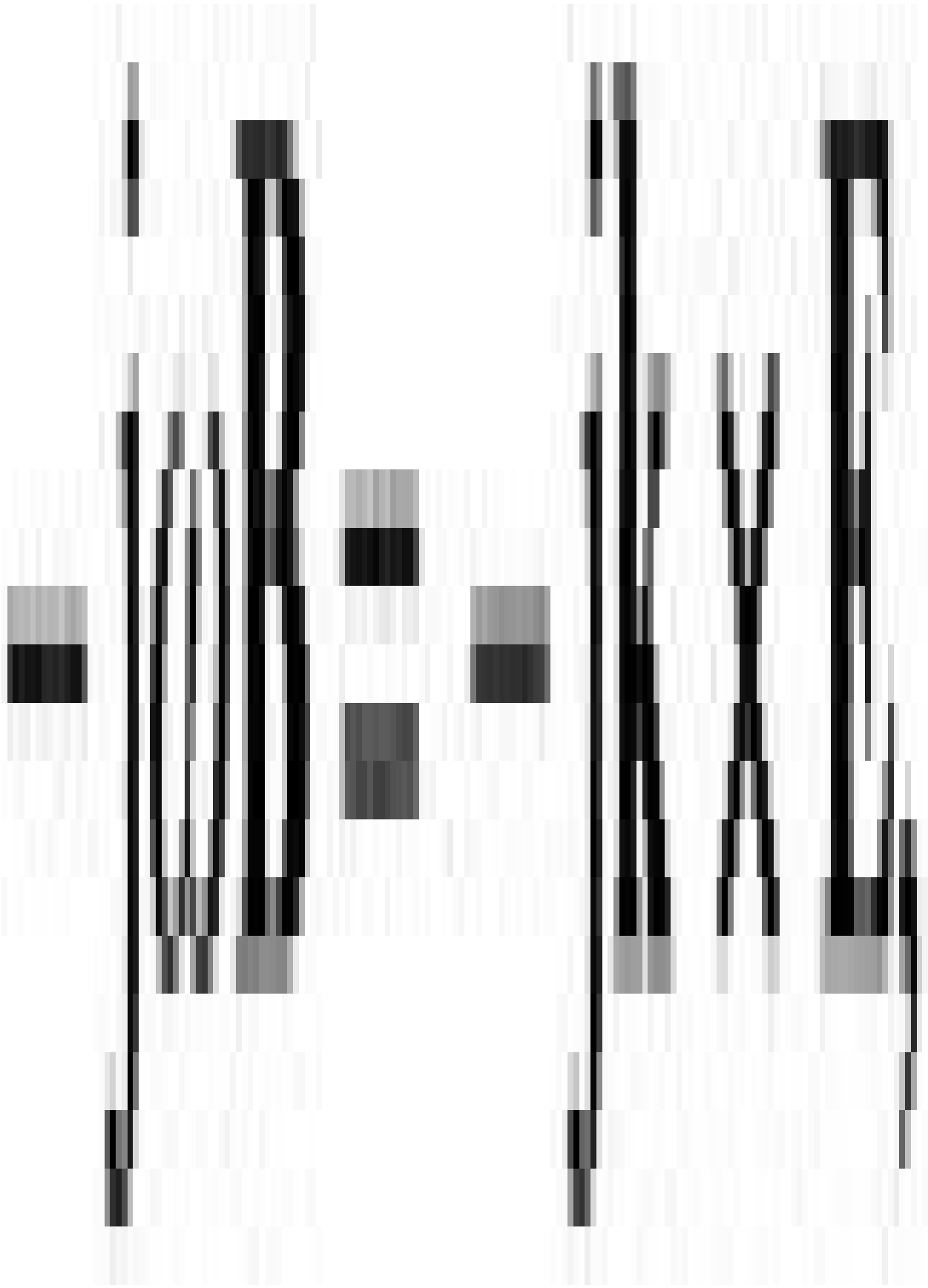
from which we see that $k \cdot E = 0$ and the vector k must be perpendicular to the vector E0. The general expression for the magnetic field B in a linearly polarized plane wave with propagation vector k is



The relationship between the fields E and B can be obtained in a straightforward way using this expression and Equation (11.18). Taking the partial derivative of the field B with respect to time and equating it to the curl of the electric field gives



Evaluation of curl E results in the equation



and hence

(11.19)

where is a unit vector in the direction of propagation of the wave. Figure 11.5 illustrates the relative directions of the electric and magnetic fields in a linearly polarized plane wave.

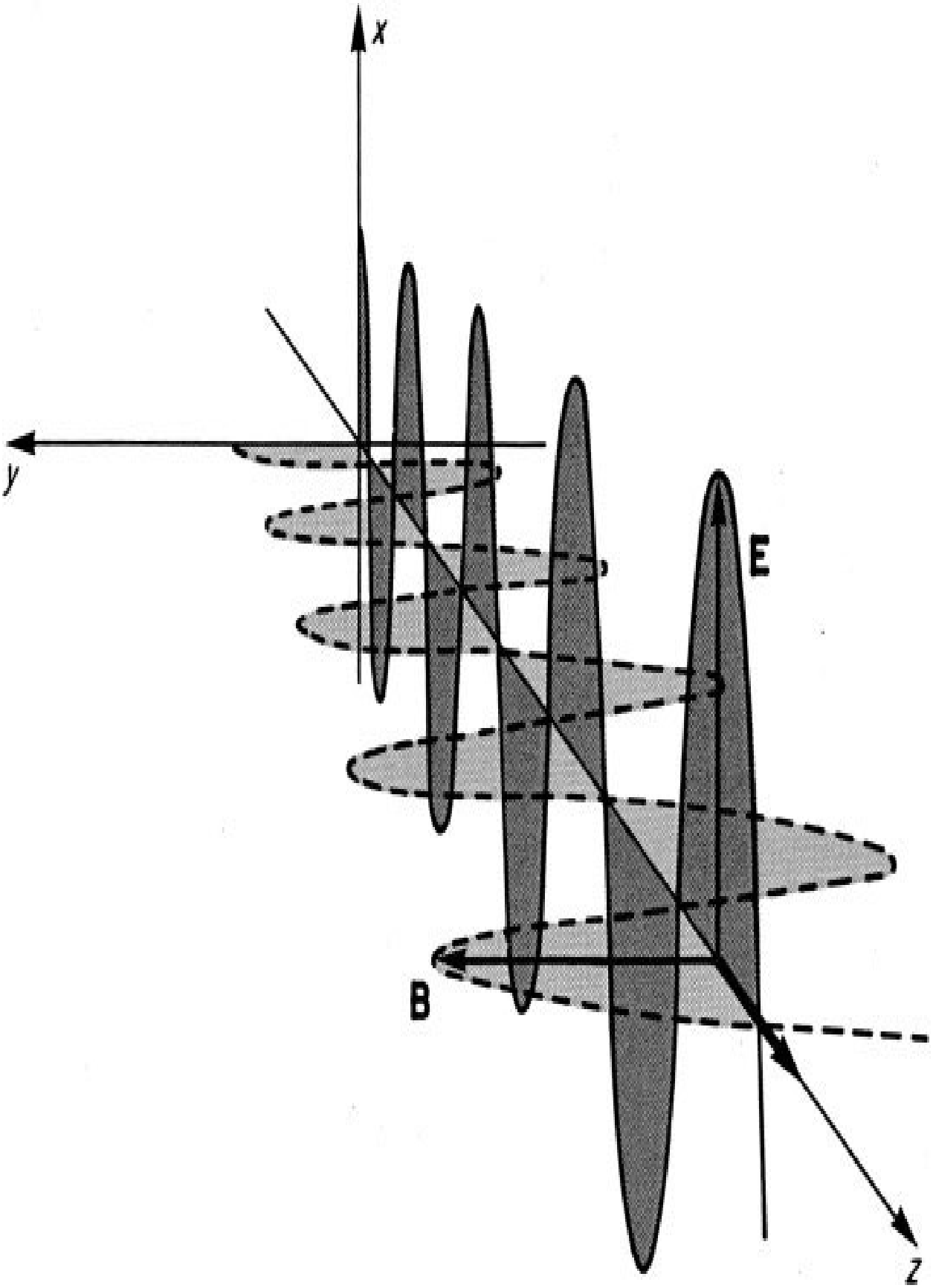
The amplitude of the magnetic field in a plane wave in free space is usually quite small. The ratio of the electric field to the magnetic field is equal to the speed of light in free space. If the amplitude of the electric field in a plane wave is 0.1 V m−1, as might be encountered in a radio wave some kilometres from a transmitter, the amplitude of the magnetic field is $0.1 \times (3 \times 10^8)^{-1}$ or 3.3×10^{-10} T.

11.2.2 Plane waves in isotropic insulating media

Let us now consider waves travelling in an isotropic, non-conducting medium of relative permittivity ε and relative permeability μ. The wave equations for the fields E and B can be derived from Maxwell's equations in a manner similar to that given in section 11.1. They are

(11.20)

Figure 11.5. The relative directions of the electric and magnetic fields in a plane wave. The wave is linearly polarized and the electric field vector oscillates along the x-axis. The wave propagates in the z-direction. The magnetic field in the wave is along the y-axis and in phase with the electric field.


x
y
E
B
z

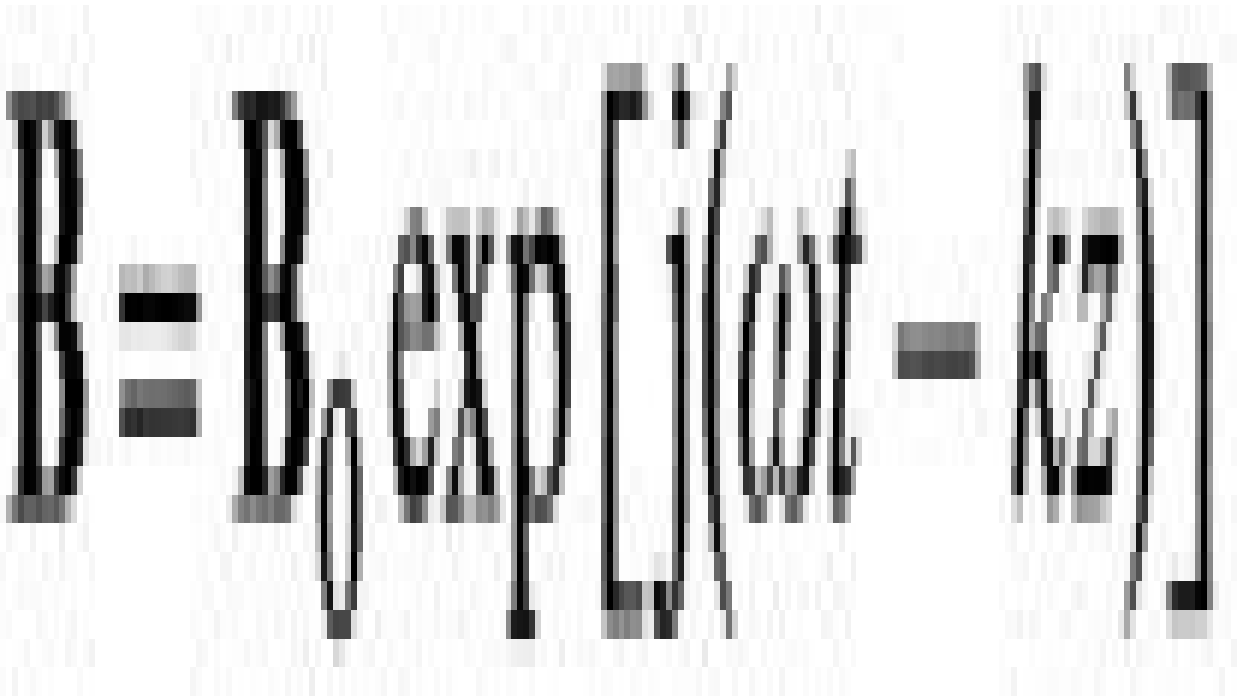
and

(11.21)

The solutions to these equations represent waves travelling with a speed υ given by

(11.22)

The monochromatic, linearly polarized, plane wave solutions to Equations (11.20) and (11.21), for a wave travelling in the positive z-direction, are



$$B = B_0 \exp[j(\omega t - kz)]$$

$$E = E_0 \exp[j(\omega t - kz)],$$

These are the same expressions as for a wave in free space. Now however, for a wave of angular frequency ω, the wavenumber k is given by

(11.23)

instead of $k = \omega/c$. In terms of the wavelength λ and the frequency v, Equation (11.23) can be written

(11.24)

The wavelength of a wave of fixed frequency is thus different when the wave is travelling in a medium to the wavelength when the wave is travelling in free space. If the product of the relative permittivity and the relative permeability of the medium is greater than unity, as is usually the case, the speed v of the wave in the medium is less than the speed c in free space, from Equation (11.22). The wavelength λ of the wave is thus decreased, as can be seen from Equation (11.24).

By arguments similar to those used in section 11.2.1 above, it can be shown that the electric and magnetic fields in plane waves travelling in isotropic media must also lie in directions perpendicular to the direction of propagation of the wave. If the direction of propagation is given by the unit vector it may also be shown in the same way that the relationship between the electric and magnetic fields in the wave in the medium is

(11.25)

The wavelength of a wave of fixed frequency changes as we go from one medium to another, as we have seen. Equation (11.22) shows that the speed of the wave in a medium depends on the frequency, since the quantities ε and μ vary with frequency. The ratio of the speed of an electromagnetic wave in free space to its speed in a medium is equal to the refractive index n of the medium for radiation of the frequency of the wave:

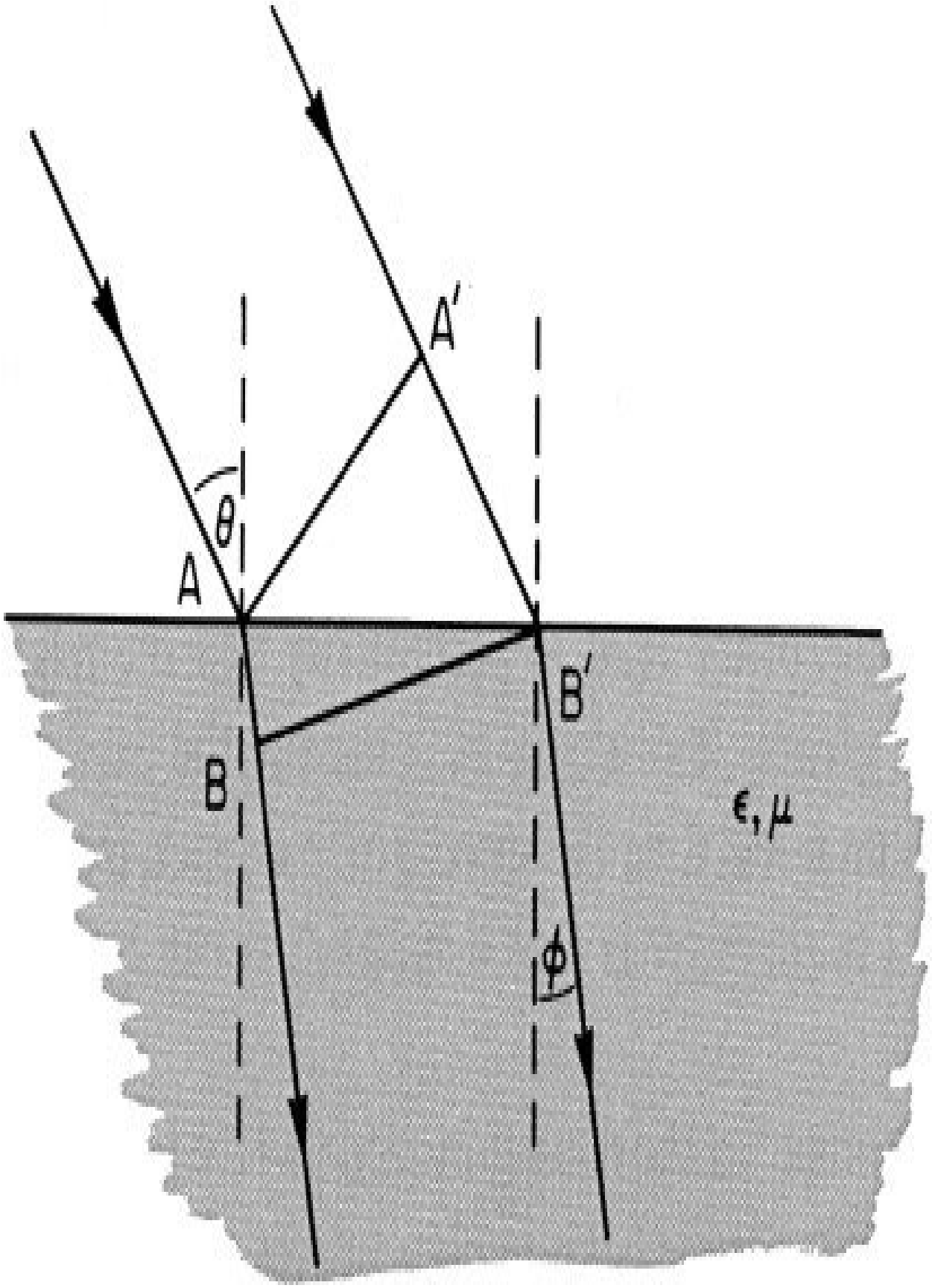
(11.26)

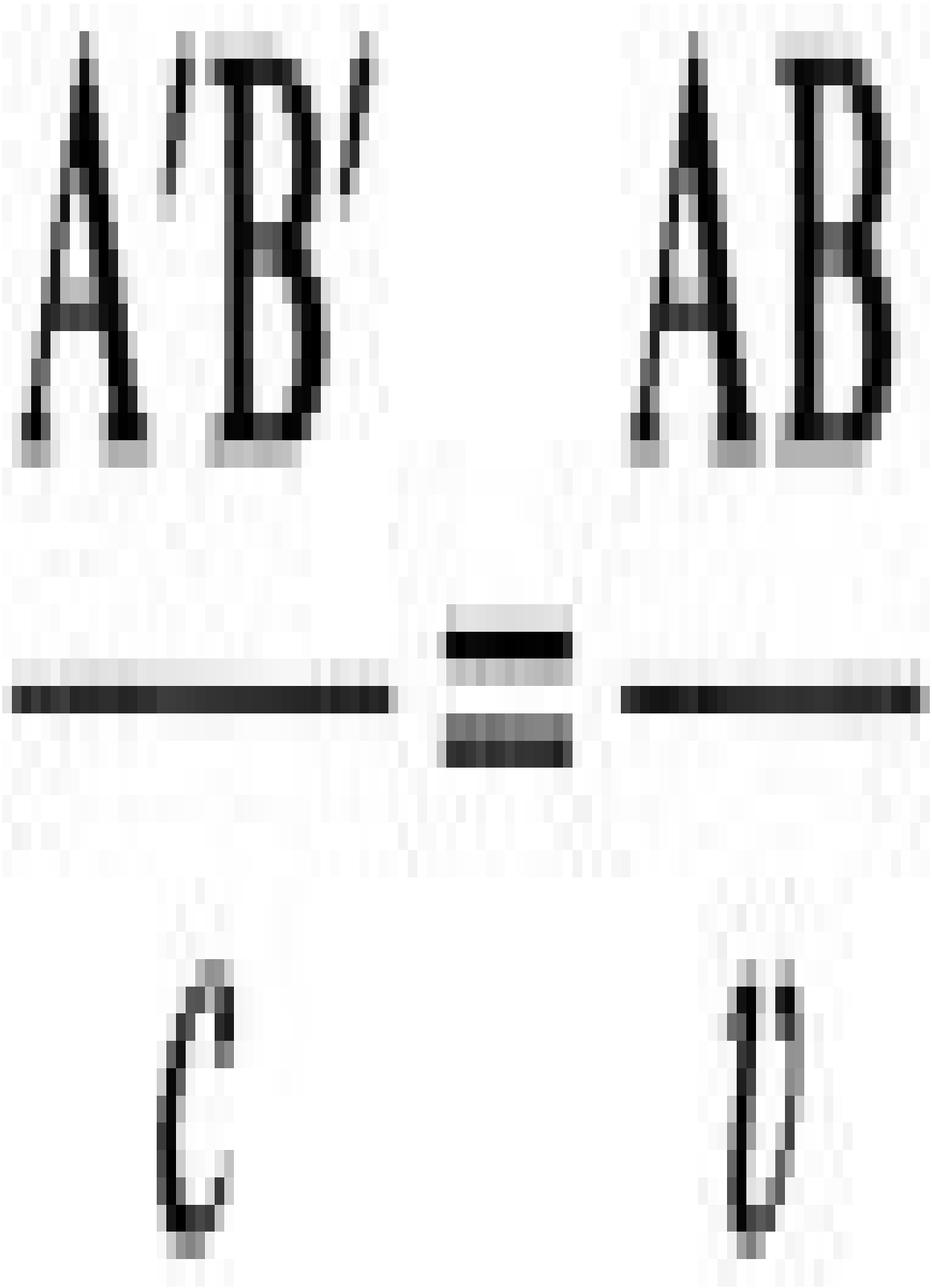
Thus the refractive index varies with frequency.

In order to show that the definition (11.26) of refractive index is the same as the definition in terms of angles of incidence and refraction of a wave at a plane boundary, consider a plane wave incident on the flat interface shown in Figure 11.6. Let the wave be incident in free space onto a medium with relative permittivity ε and relative permeability μ. The wave front of the incident wave is AA′ and the wave front of the refracted wave is BB′. Along the surfaces AA′ and BB′ the phase of the wave is constant, hence the change in phase over the distance AB equals the change over A′B′. The frequency of the wave does not change as it goes into the medium. Any changes in the properties of the wave are due to the oscillations of the electrons in the atoms or molecules of the medium as they interact with the incident field. The electrons in turn radiate a field which modifies the original wave, but they radiate at the same frequency and so the final product is a wave of the original frequency. The wavelength and speed of the wave change but the frequency does not. Hence the phases at the points B and B′ are the same if the times taken for the radiation to go from A to B and from A′ to B′ are the same. For this to be so we require that

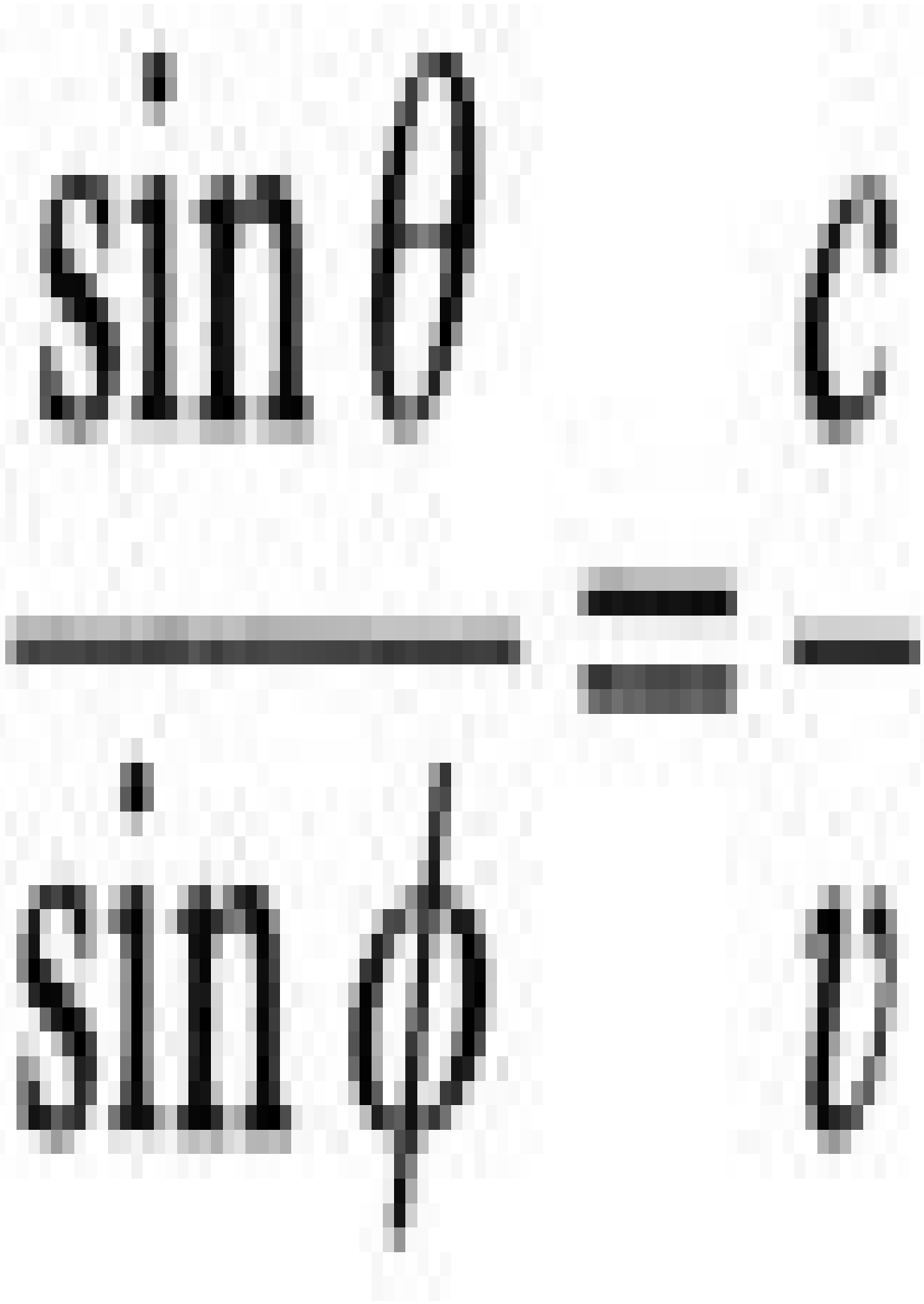
Figure 11.6. A plane wave refracted at the plane boundary between two media. The directions of propagation of the wave in air and in the medium are given by

the directions of the rays A′B′ and AB respectively.


A′
θ
A
B′
B
ϵ, μ
ϕ

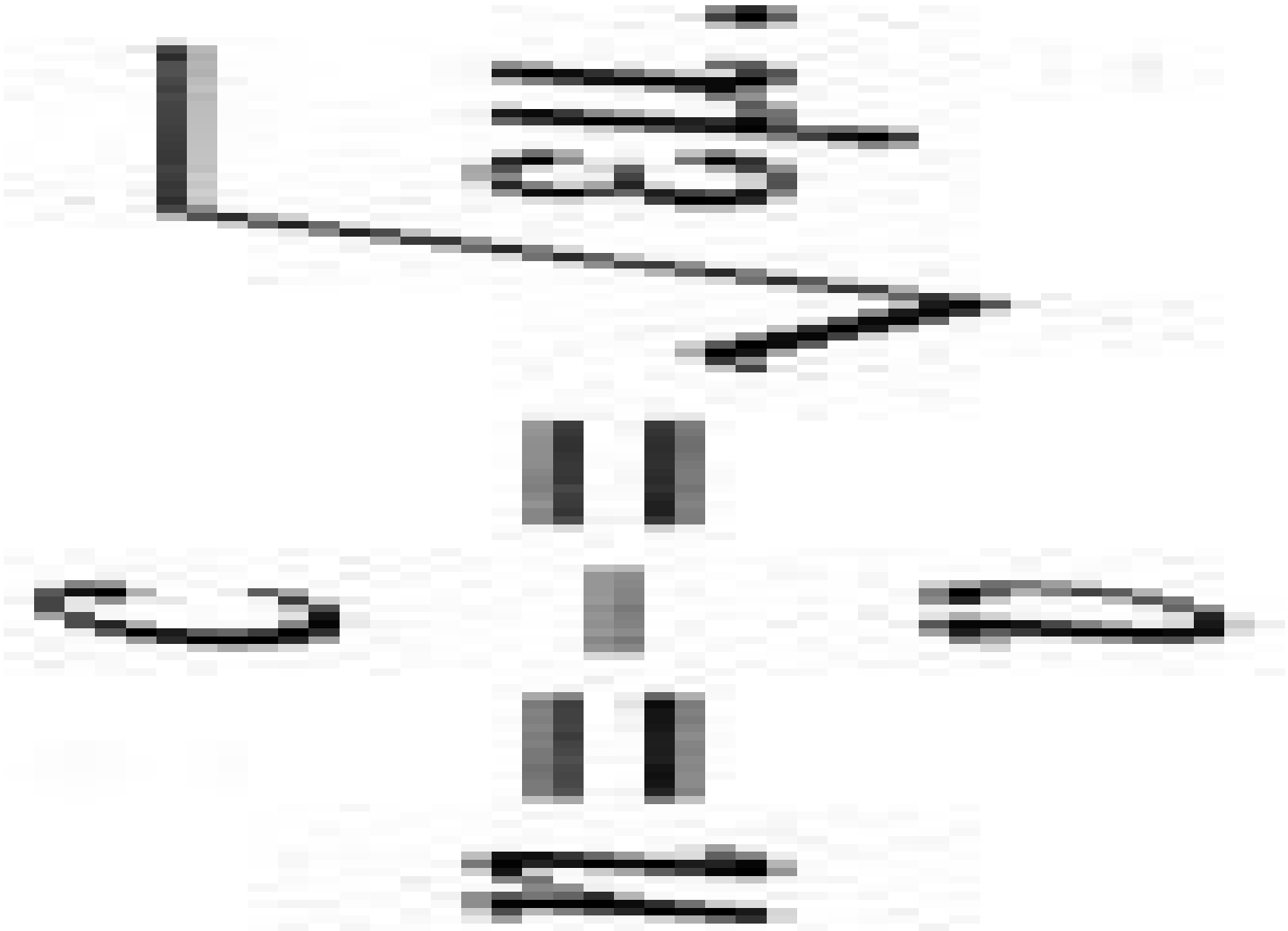


or



But relating the angles θ and ϕ through the familiar Snell's law, we have

and hence



Usually it is a good enough approximation to put the relative permeability equal to unity and write

(11.27)

If the relative permittivity of a substance is measured from zero frequency up to visible light frequencies it is observed to fall off slowly as the frequency increases. However, at the higher frequencies, there exist anomalies superimposed on the slow fall off which occurs over small frequency ranges. Let us take water as an example. In static fields and in alternating fields at low frequencies the relative permittivity of water is about 80 at room temperature. As the frequency of the alternating electric field increases above 100 MHz, the relative permittivity begins to decrease. It falls throughout the far infrared region of the spectrum, and is down to about 8 at 10 GHz. At low frequencies the chief contribution to the polarization arises from the preferred orientation along the field of the permanent dipole moment of the water molecules. As the frequency increases, the molecules, which are restricted by frictional torques from other nearby molecules, are unable to turn fast enough to follow the rapid variations of the field direction. The contribution to the relative permittivity from the permanent dipole moments thus falls off to zero. As the frequency increases further, into the near infrared and visible parts of the spectrum, the induced dipole moment of each molecule in turn fails to keep up with the field alternations, and the relative permittivity falls even further. The anomalies in the curve showing how the relative permittivity varies with frequency occur at frequencies corresponding to the internal modes of excitation of the water molecule. At a frequency of 5.1×10^{15} Hz (yellow light from a sodium vapour lamp) the refractive index of water at room temperature is down to 1.33, corresponding to a relative permittivity of 1.77.

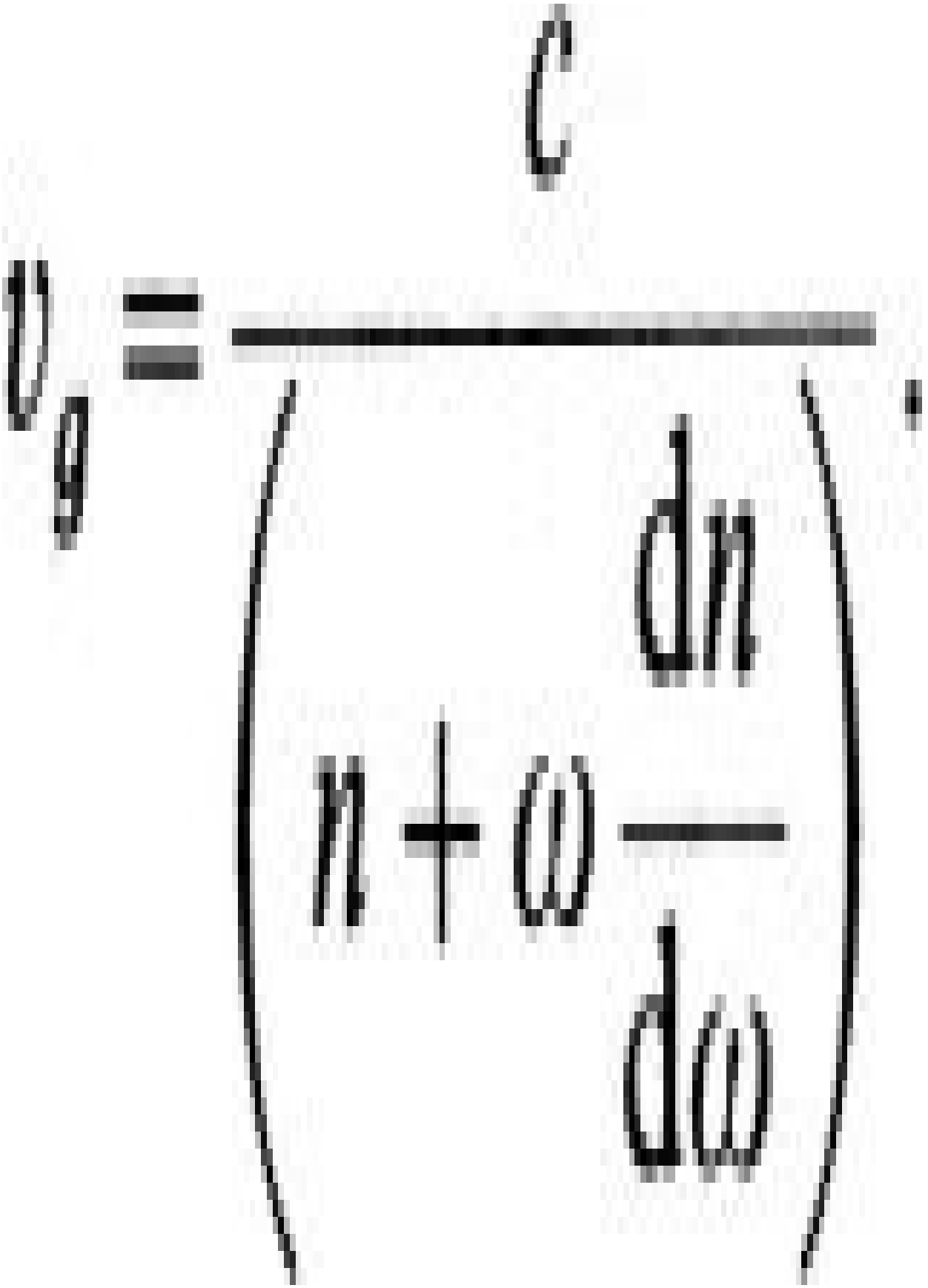
11.3. DISPERSION

The variation of refractive index with frequency arises because waves of different frequency move through media at different speeds. This phenomenon is known as dispersion and will be discussed in this section. Dispersion is responsible for the colours seen when a beam of white light is split into its component wavelengths after being refracted through a prism. Different wavelengths move at unequal speeds in the glass and hence have different refractive indices and are bent by different amounts.

Pure monochromatic waves do not exist in nature. The reason we have been discussing monochromatic waves (and why we will continue to do so in following sections) is that they are mathematically quite simple and serve to illustrate many of the important properties of electromagnetic wave propagation. Waves that do actually occur are made up of monochromatic waves with several frequencies, and form what is known as a wave group. If the wave group is nearly monochromatic the frequencies of the group are clustered in a narrow region around the main frequency. If a wave group is propagating in a dispersive medium each frequency component travels with a slightly different speed. The speed υ with which a component with angular frequency ω travels is given by Equation (11.22), and the wavenumber k of the component wave by Equation (11.23). This speed is called the wave velocity or phase velocity of that component of the group which has angular frequency ω. The energy in the group of waves travels with the group velocity υg. This depends on the dispersion of the different frequencies in the group, i.e. on the variation of the speed of the waves with frequency, and can be shown* to be given by

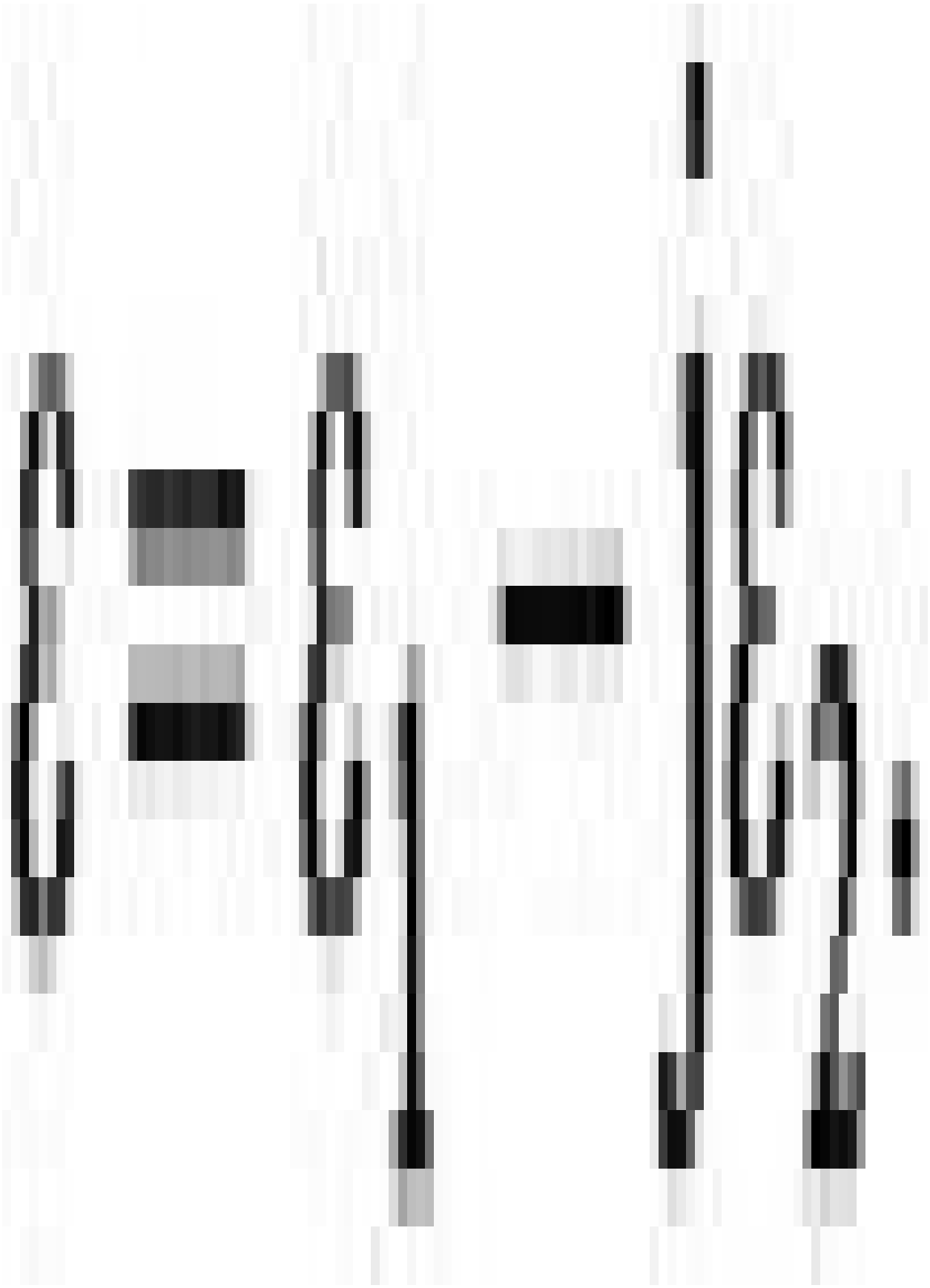
(11.28)

which in turn, for electromagnetic waves, becomes

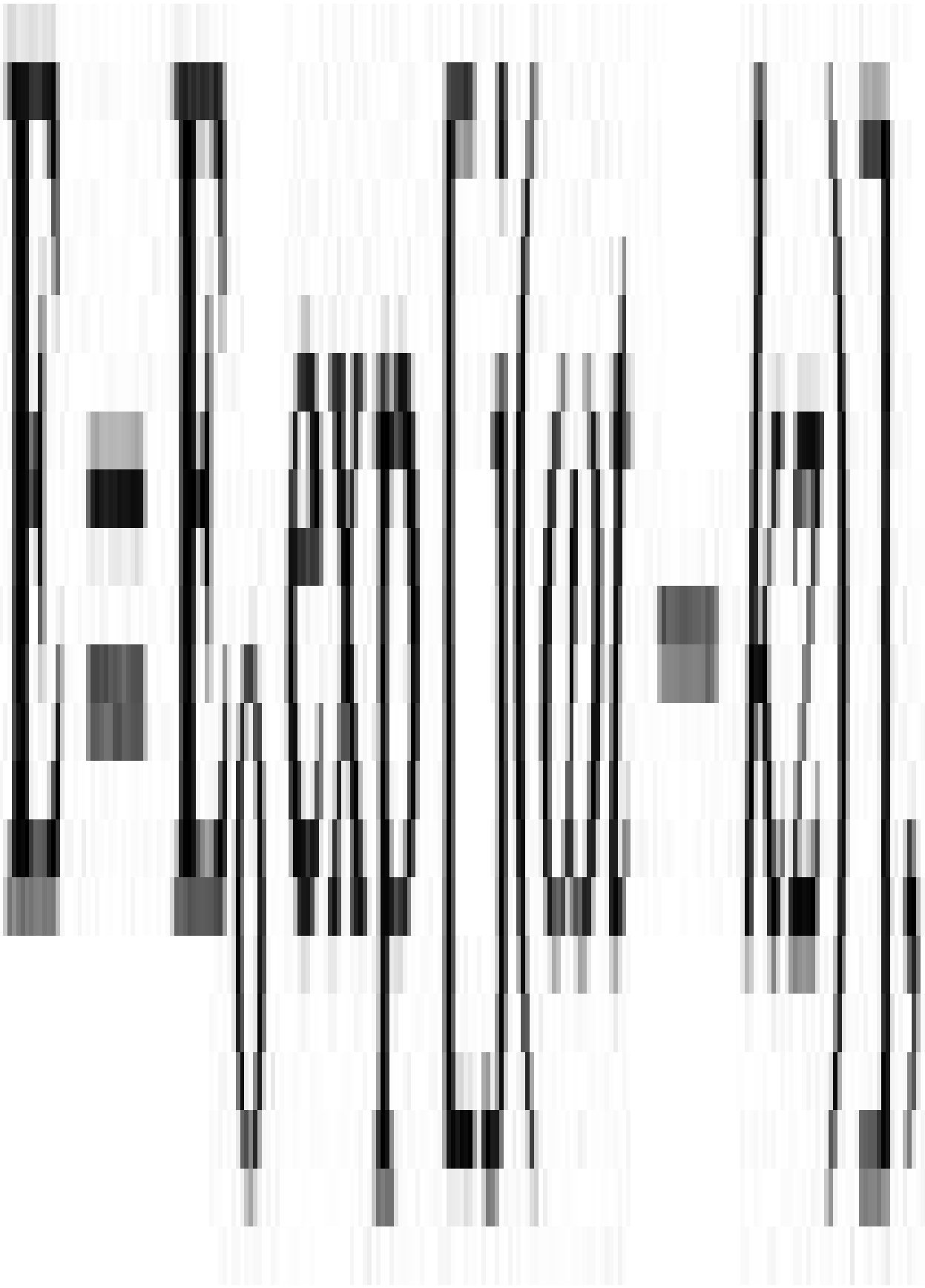


In some situations wave velocities can exceed the speed of light in free space. An example is in the ionosphere, which has a refractive index less than unity for high frequency waves (see Problem 11.9) and so can totally reflect radio waves incident at a sufficiently oblique angle. There is no contradiction of the special theory of relativity since the energy of a wave travelling through the ionosphere moves with the group velocity, which is always less than the speed of light in free space.

As discussed in section 11.2.2, using the example of water, materials usually show a smooth decrease of relative permittivity with increasing frequency. Superimposed on the smooth fall-off, at high frequencies, are anomalies which occur at the natural resonant frequencies of internal modes of excitation of the material's microscopic constituents. Close to the resonant frequencies energy is absorbed from the incident wave and transferred to internal excitation. Energy can also be absorbed at frequencies far from resonances and dissipated as heat via frictional processes, as described earlier for water. Absorption of energy from an electromagnetic wave results in the field amplitudes becoming smaller, since these amplitudes determine the wave's energy. Using complex notation for the fields in the wave, absorption of energy can be described by attributing to the medium a complex relative permittivity ε, where



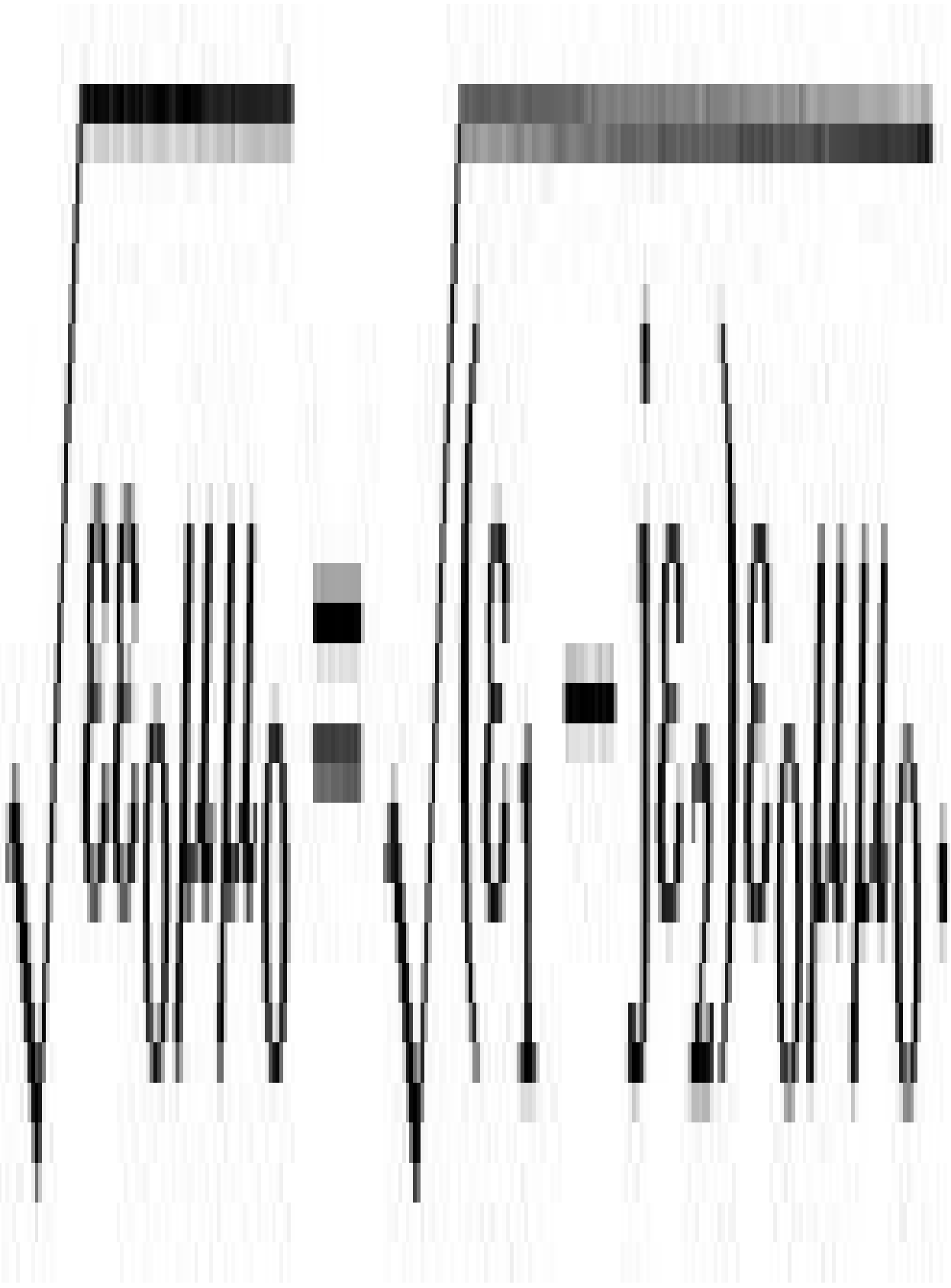
A plane wave progressing in the medium can be represented by Equation (11.16),



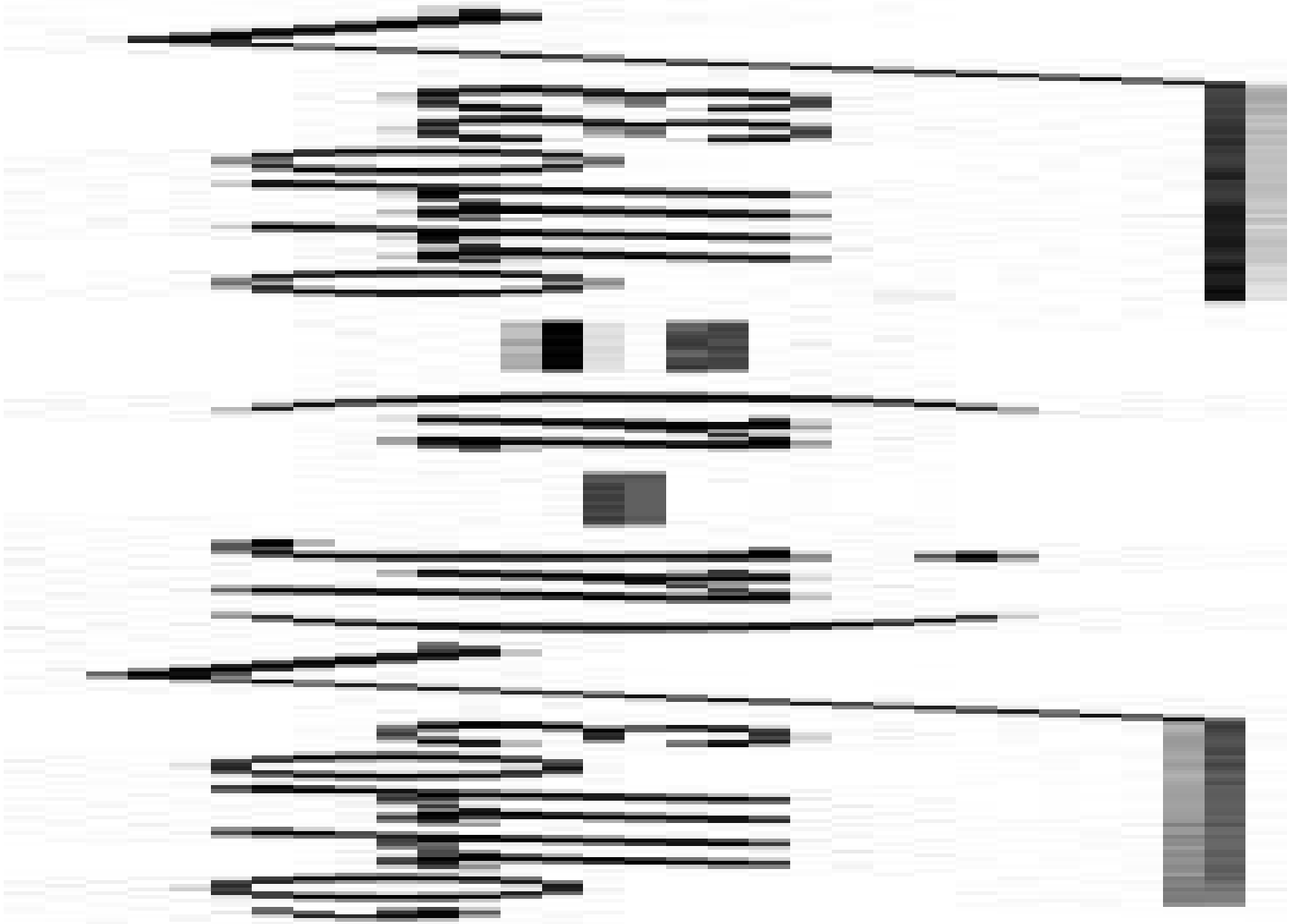
or equivalently (using Equations (11.22) and (11.23)) by

(11.29)

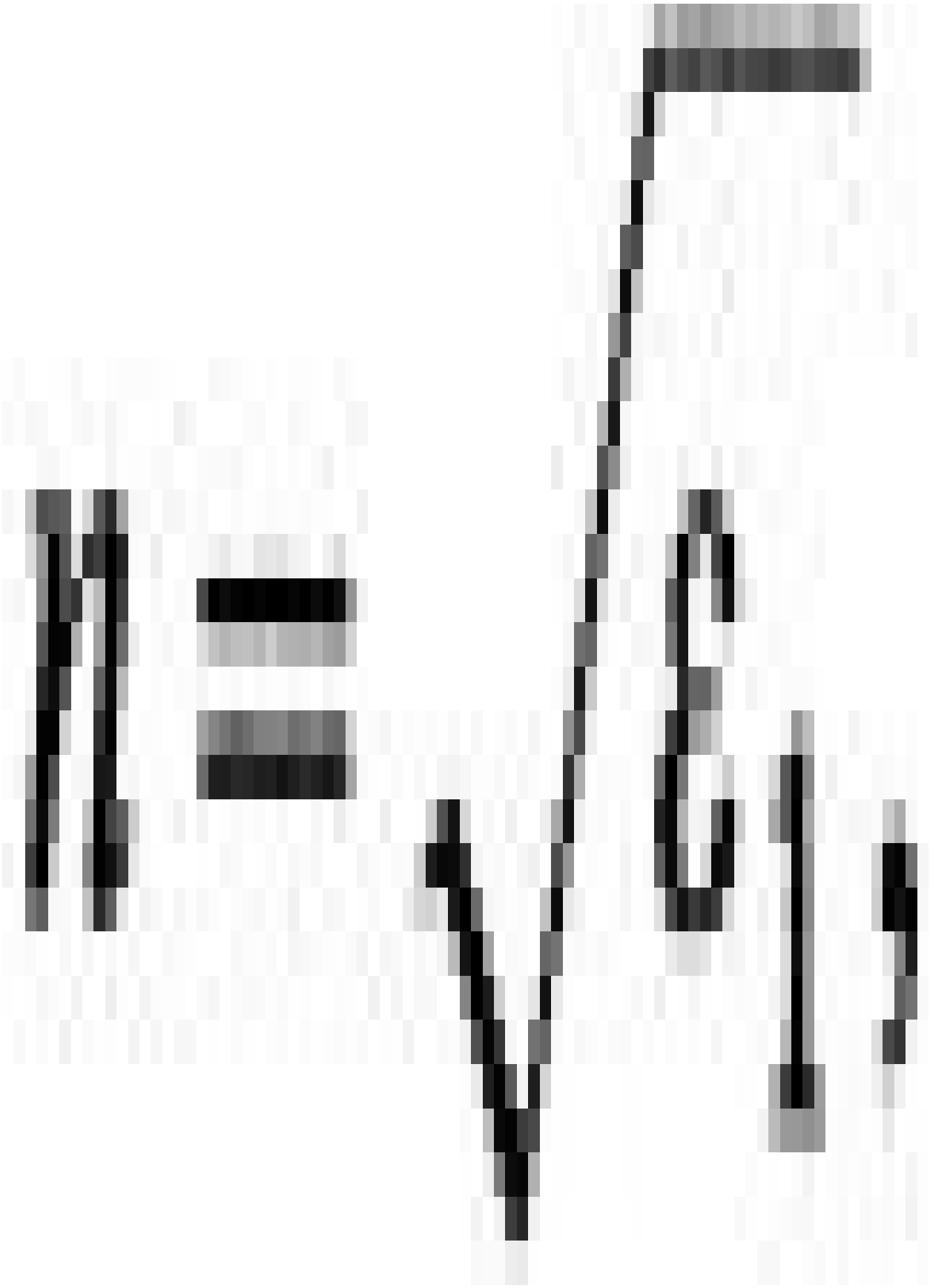
If ε is complex



This may be written in the form



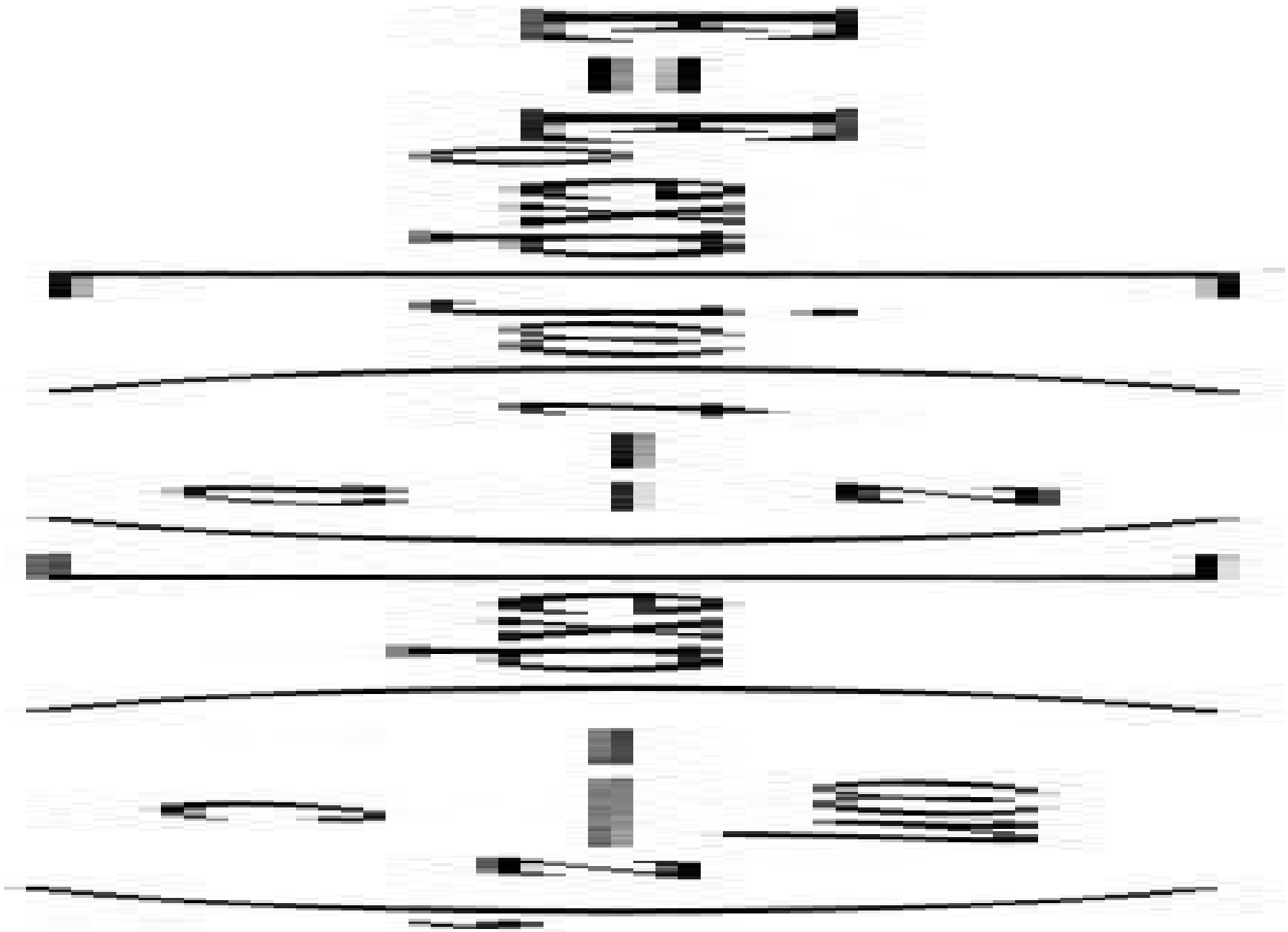
where



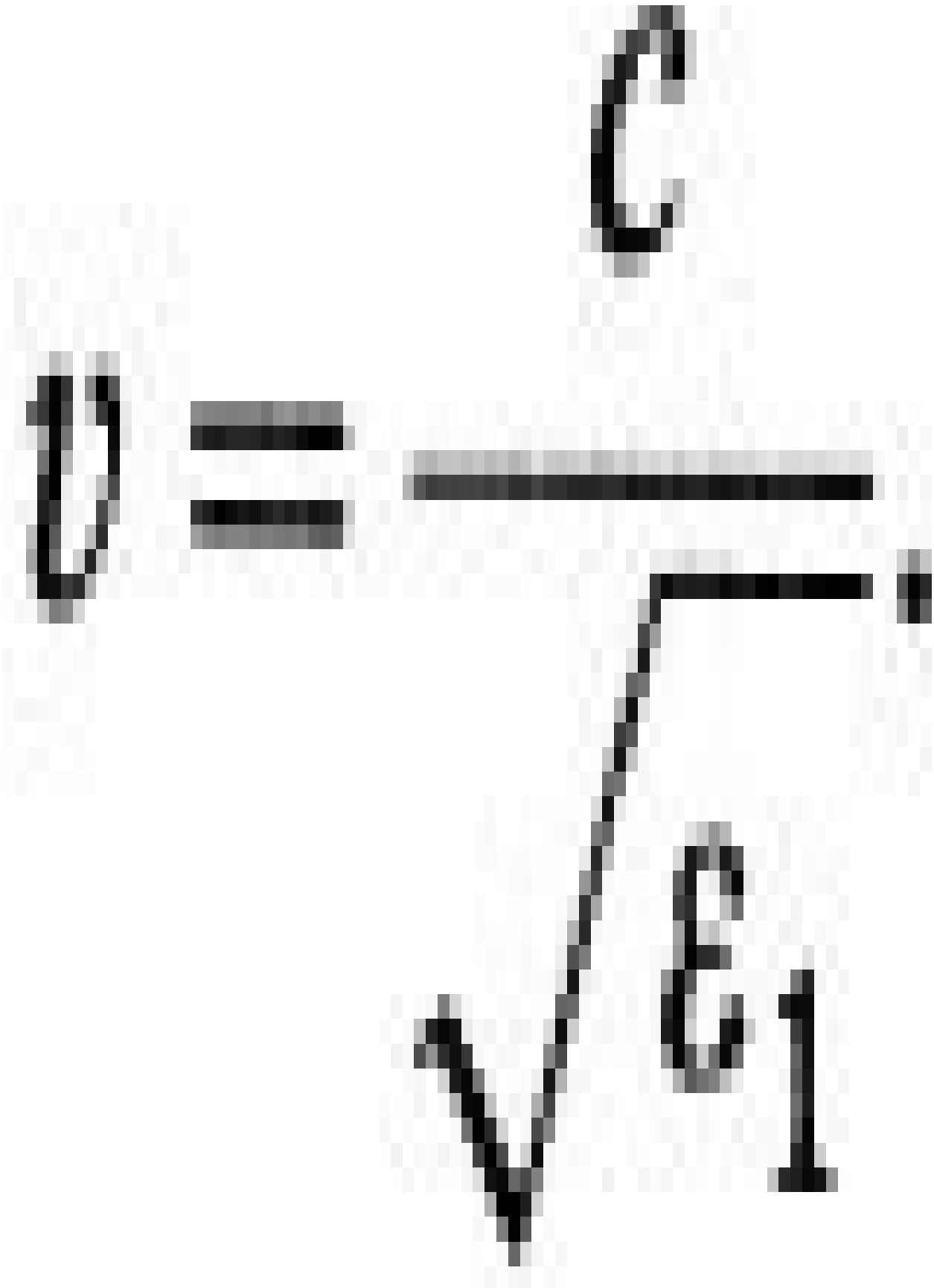
and

We have assumed that ε2 is small compared to ε1 corresponding to a small absorption of energy. If we put the relative permeability μ equal to unity, Equation (11.29) can now be rewritten

(11.30)



where

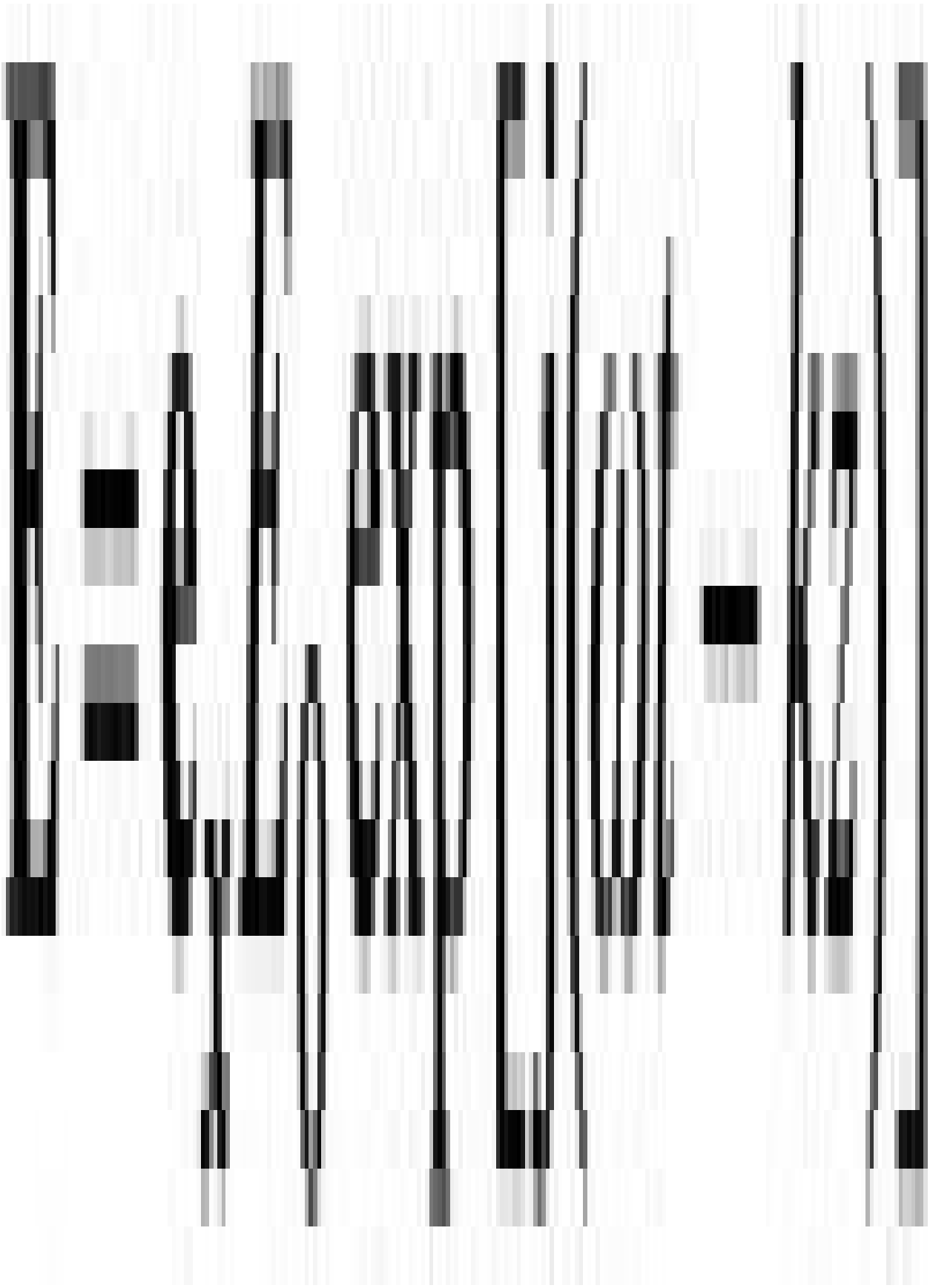


This expression represents a plane wave whose amplitude is decreasing exponentially with z. The amplitude falls to one half of its initial value over a distance equal to

The quantities n and η vary with frequency. For radiation of wavelength 5×10^{-7} m (yellow visible light) in glass, η is very small. As the frequency is increased into the ultraviolet region of the spectrum, η increases due to the excitation of the electrons in the molecules of the glass. Glass is opaque to ultra-violet wavelengths shorter than about 2×10^{-7} m. As another example most of the infrared waves with wavelengths between about 10^{-3} and 10^{-5} m cannot penetrate the atmosphere because of absorption due to the excitation of rotations in the air molecules.

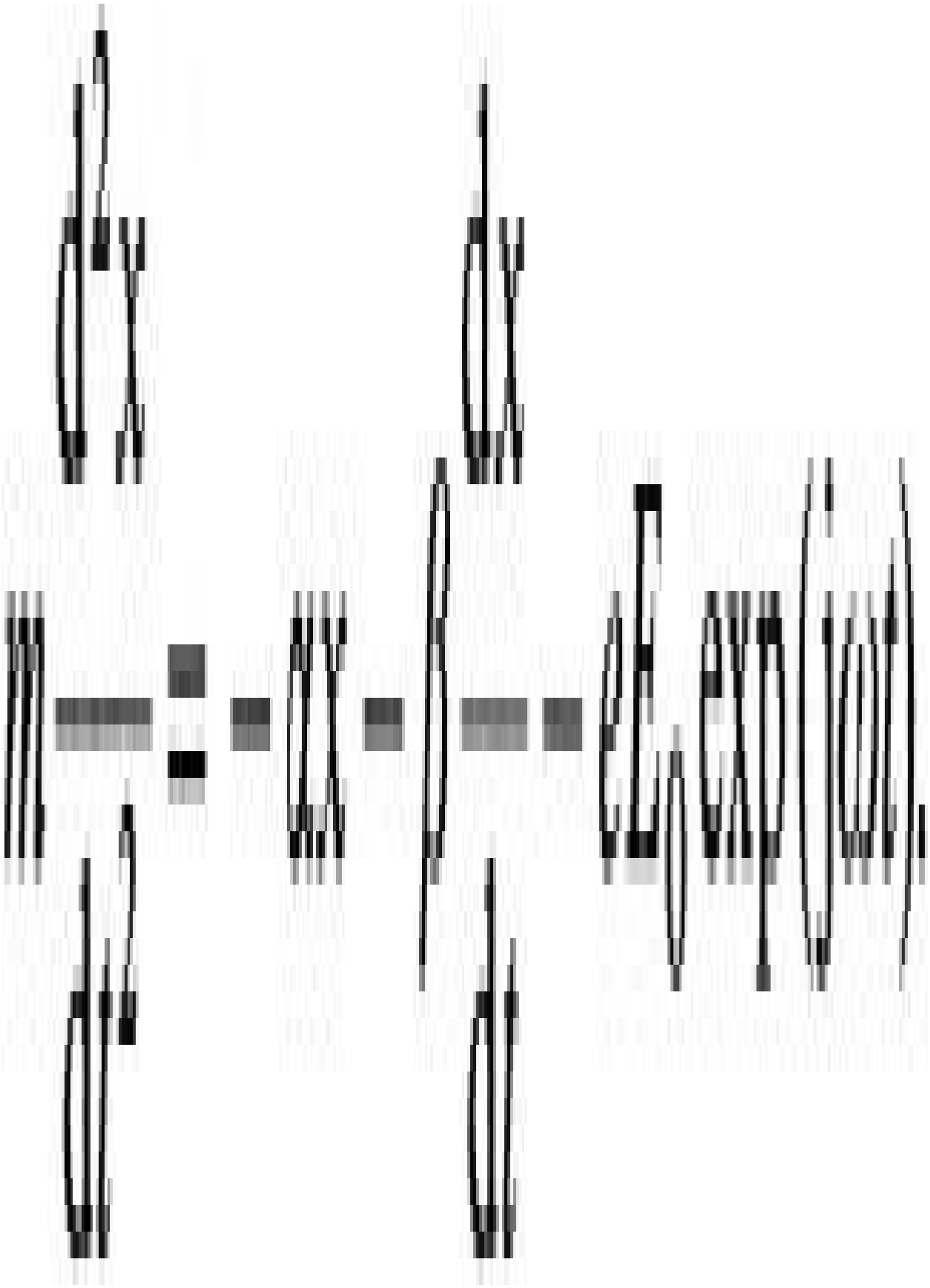
Classical theory cannot properly describe absorption between microscopic energy levels. The absorption of microwaves in the atmosphere, for example, needs the quantum theory of the interaction of the electromagnetic field with atmospheric molecules for its interpretation. However, some physical insight into the mechanism of absorption at high frequencies and into how the real and imaginary parts of the relative permittivity vary with frequency near an absorption resonance can be obtained using a simple classical model.

Consider a molecule in which an electron of mass m is bound to a centre by a restoring force proportional to displacement. Let the constant of proportionality be α. In the presence of an electromagnetic wave with electric field given by

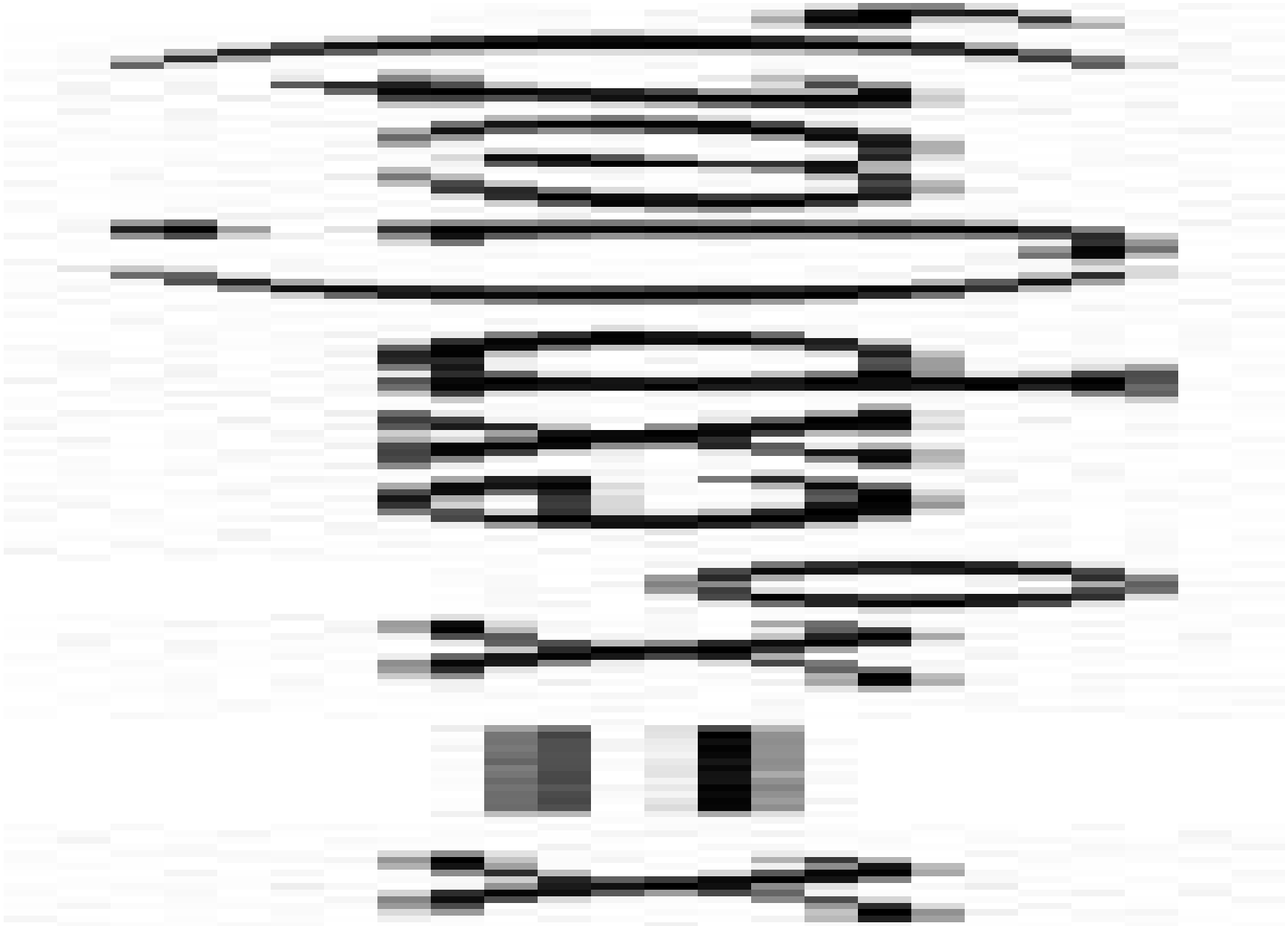


the electron undergoes forced harmonic vibrations with equation of motion

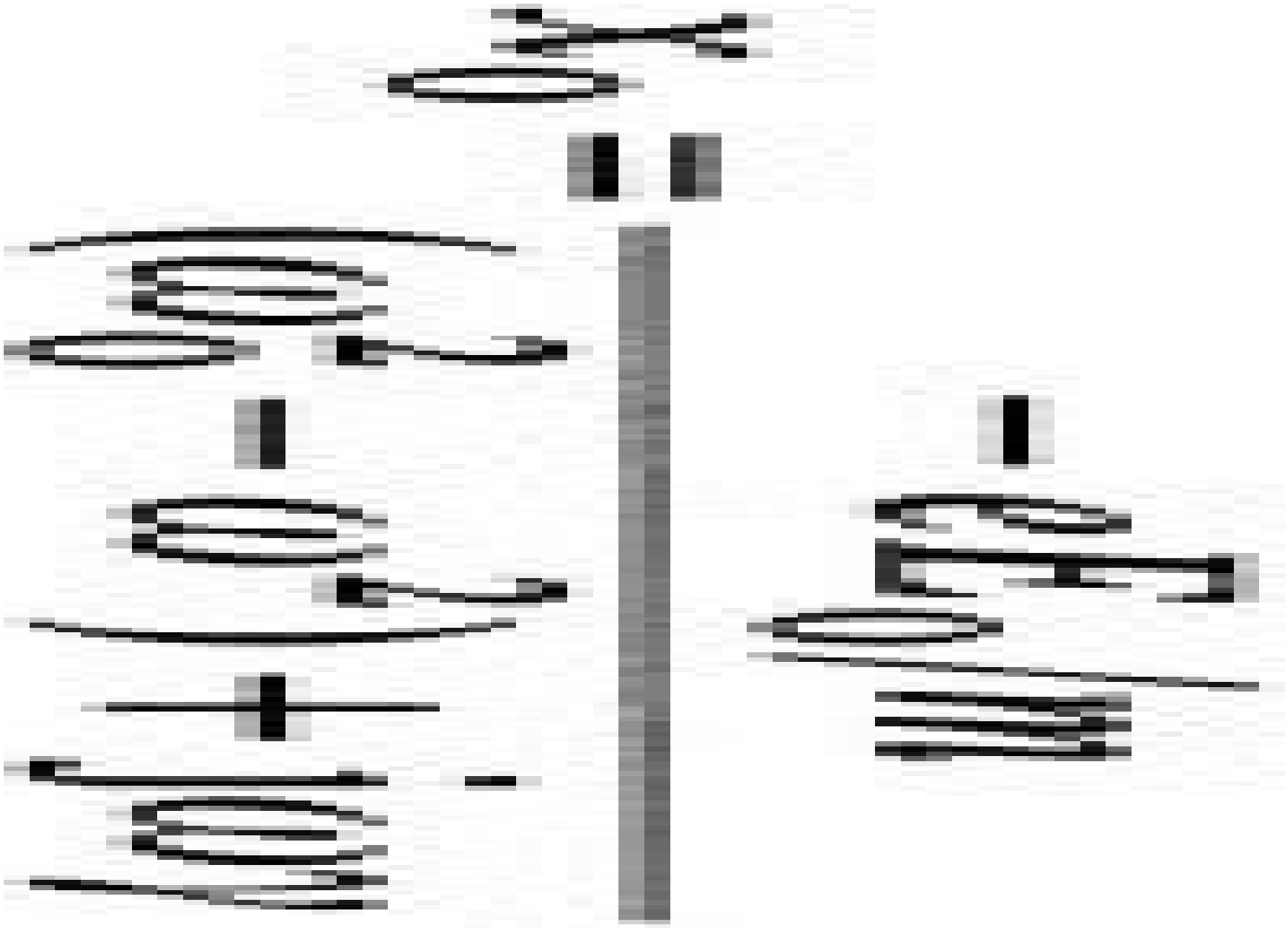
(11.31)



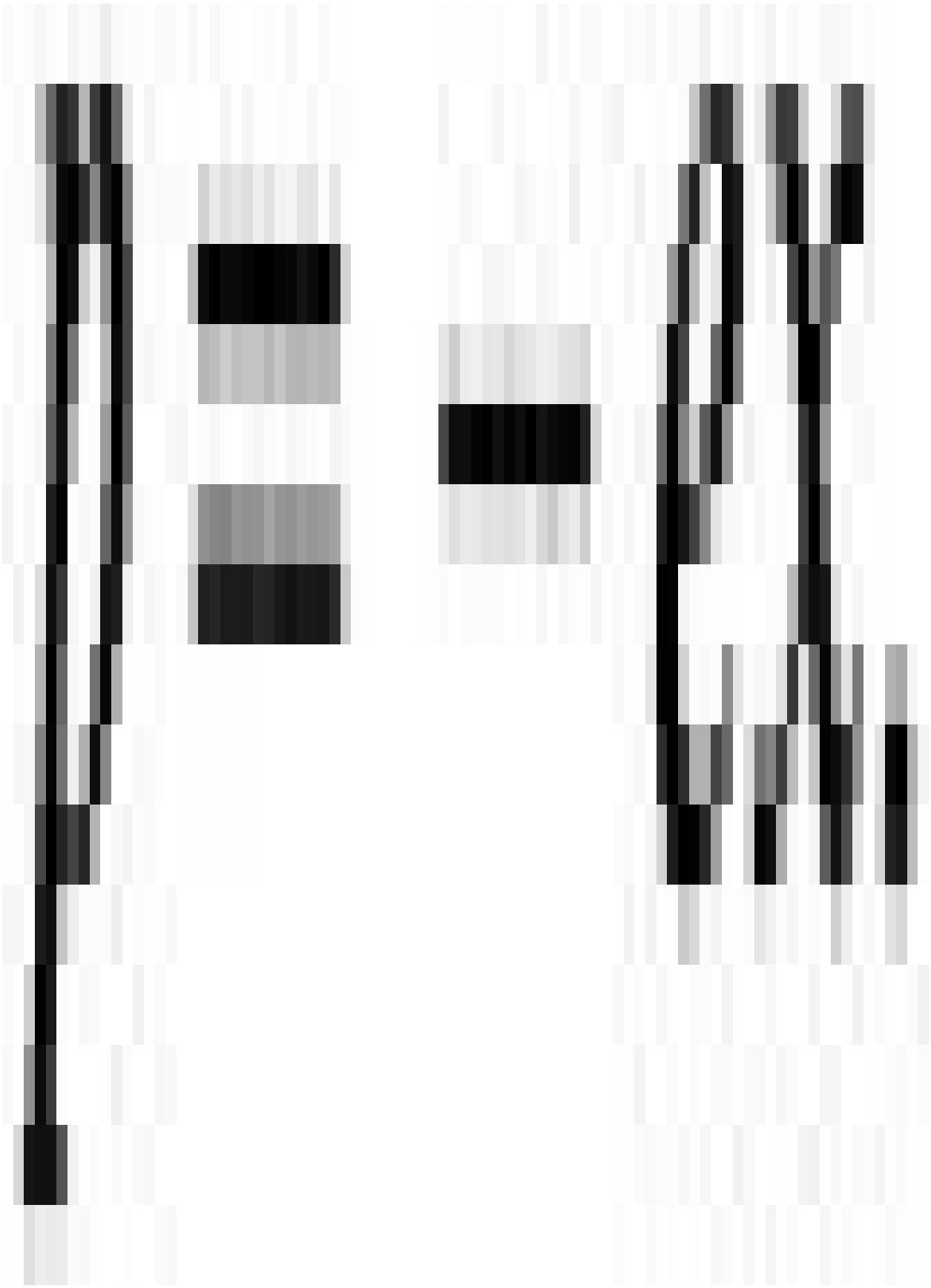
In this equation we have introduced a dissipative force proportional to velocity, with constant of proportionality β, and have chosen the z-coordinate of the electron to be $z = 0$. The solution to Equation (11.31) is



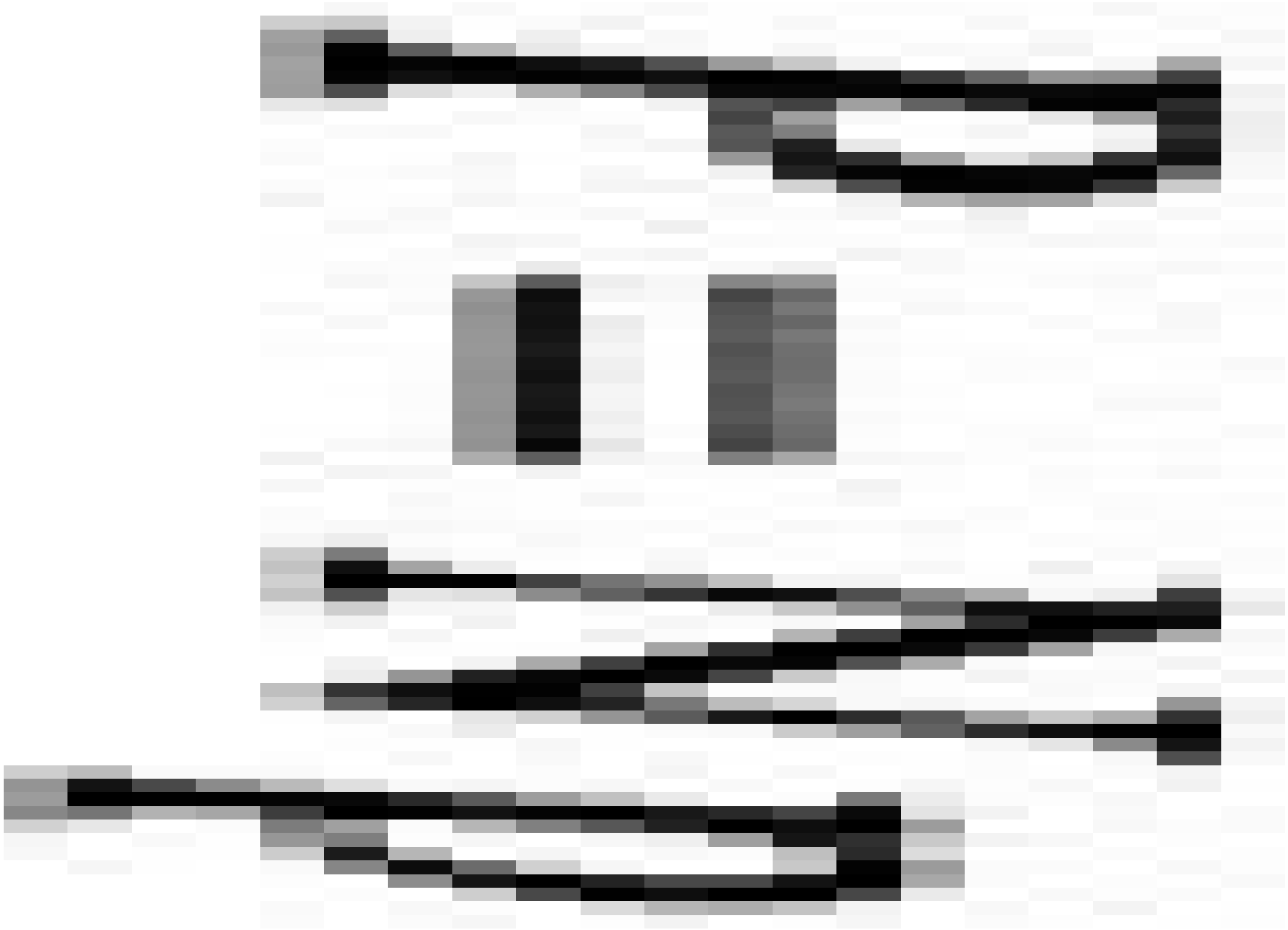
with complex x0 given by



where is the resonant angular frequency of the electron bound to its centre, and the parameter γ is equal to β/m. The dipole moment of the electron is in the x-direction, and has size



Assuming that the only contribution to the polarization P of the medium comes from electrons similar to the one we are considering, of which there are N per unit volume, then from Equation (2.4)



and the electric susceptibility χ_E is given by Equation (2.6) to be

$$\chi_E = \frac{P}{\varepsilon_0 E}$$

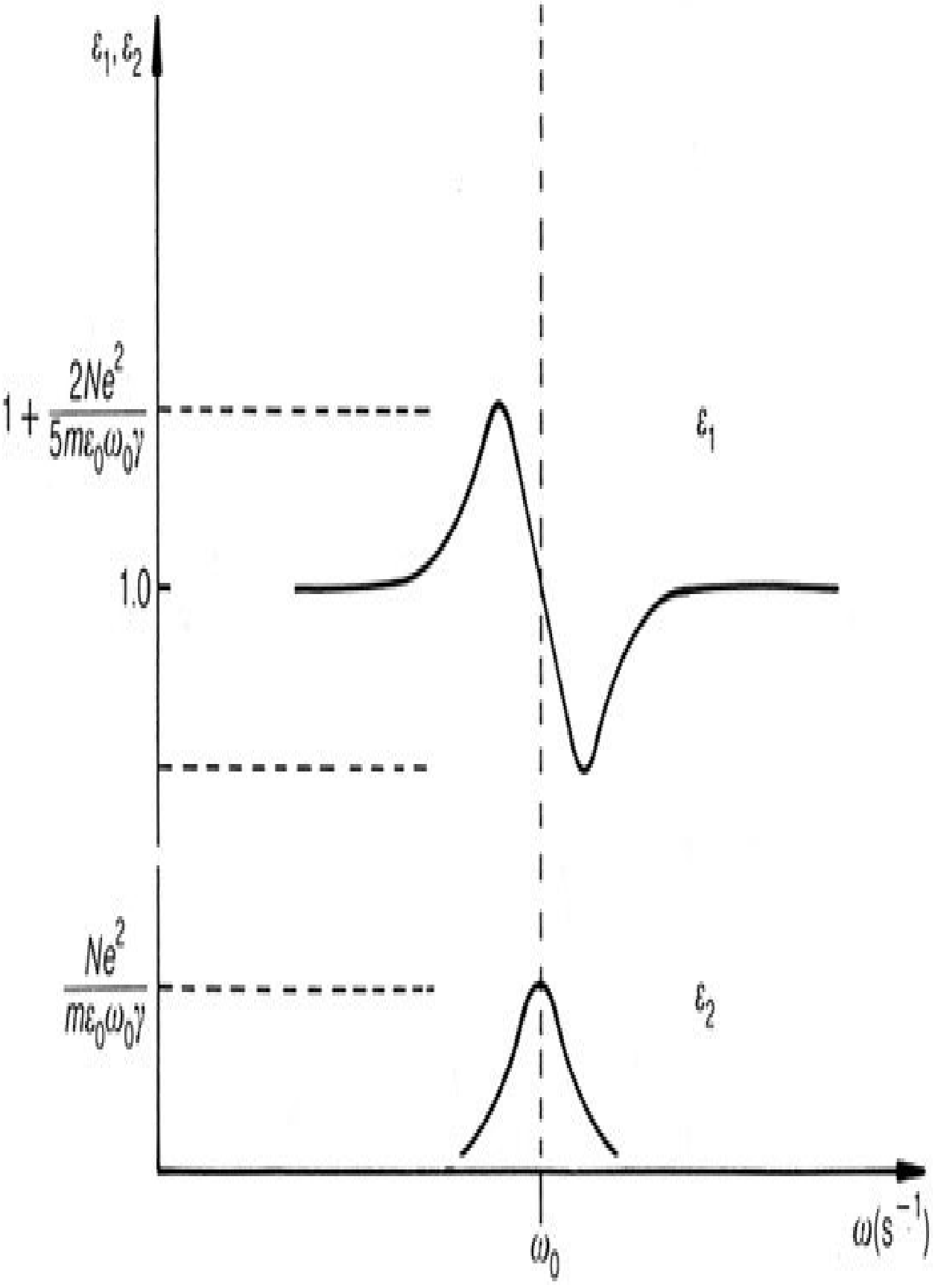
$$= \frac{Ne^2/m\varepsilon_0}{(\omega_0^2 - \omega^2) + j\omega\gamma}$$

The complex relative permittivity $\varepsilon = \varepsilon 1 - j\varepsilon 2$ is equal to $(1+\chi E)$ from Equation (2.7), from which the real and imaginary parts $\varepsilon 1$ and $\varepsilon 2$ can be determined to be

$$\varepsilon_1 = 1 + \frac{Ne^2(\omega_0^2 - \omega^2)/m\varepsilon_0}{(\omega_0^2 - \omega^2)^2 + \omega^2\gamma^2},$$

$$\varepsilon_2 = \frac{Ne^2\omega\gamma/m\varepsilon_0}{(\omega_0^2 - \omega^2)^2 + \omega^2\gamma^2}.$$

Figure 11.7. Curves showing the variation of the real and imaginary parts of the complex relative permittivity in the vicinity of an absorption resonance.


$\varepsilon_1, \varepsilon_2$
$1 + \frac{2Ne^2}{5m\varepsilon_0\omega_0\gamma}$
1.0
ε_1
$\frac{Ne^2}{m\varepsilon_0\omega_0\gamma}$
ε_2
ω_0
$\omega(s^{-1})$

These expressions are approximate owing to the simplicity of the model and because we have assumed that the electric field acting on the bound electron is the macroscopic field E in the wave, i.e. we have ignored the difference, discussed in section 2.2, between the local and macroscopic fields. However the expressions give a qualitative description of the variation of relative permittivity near an absorption resonance. The variation is shown in Figure 11.7. The frequency interval over which variations in $\varepsilon 1$ and $\varepsilon 2$ occur is characterized by the parameter γ, which is measured in s^{-1}. This parameter can vary over a wide range but a typical value is about $10^8\ s^{-1}$ for a resonance encountered at optical frequencies. Smaller values are met at lower frequencies. Quantum mechanically the reciprocal of γ corresponds to the lifetime of the allowed atomic level, at an excitation energy of $\omega 0$, which corresponds to the resonance. The symbol represents Planck's constant divided by 2π, first discussed in this book in section 4.4.3. A real atom has many resonances at different frequencies, corresponding to the excited quantum states which can be reached by absorption from the ground state, and a more realistic expression for the electric susceptibility is given by summing over all of them.

11.4. ENERGY IN ELECTROMAGNETIC WAVES

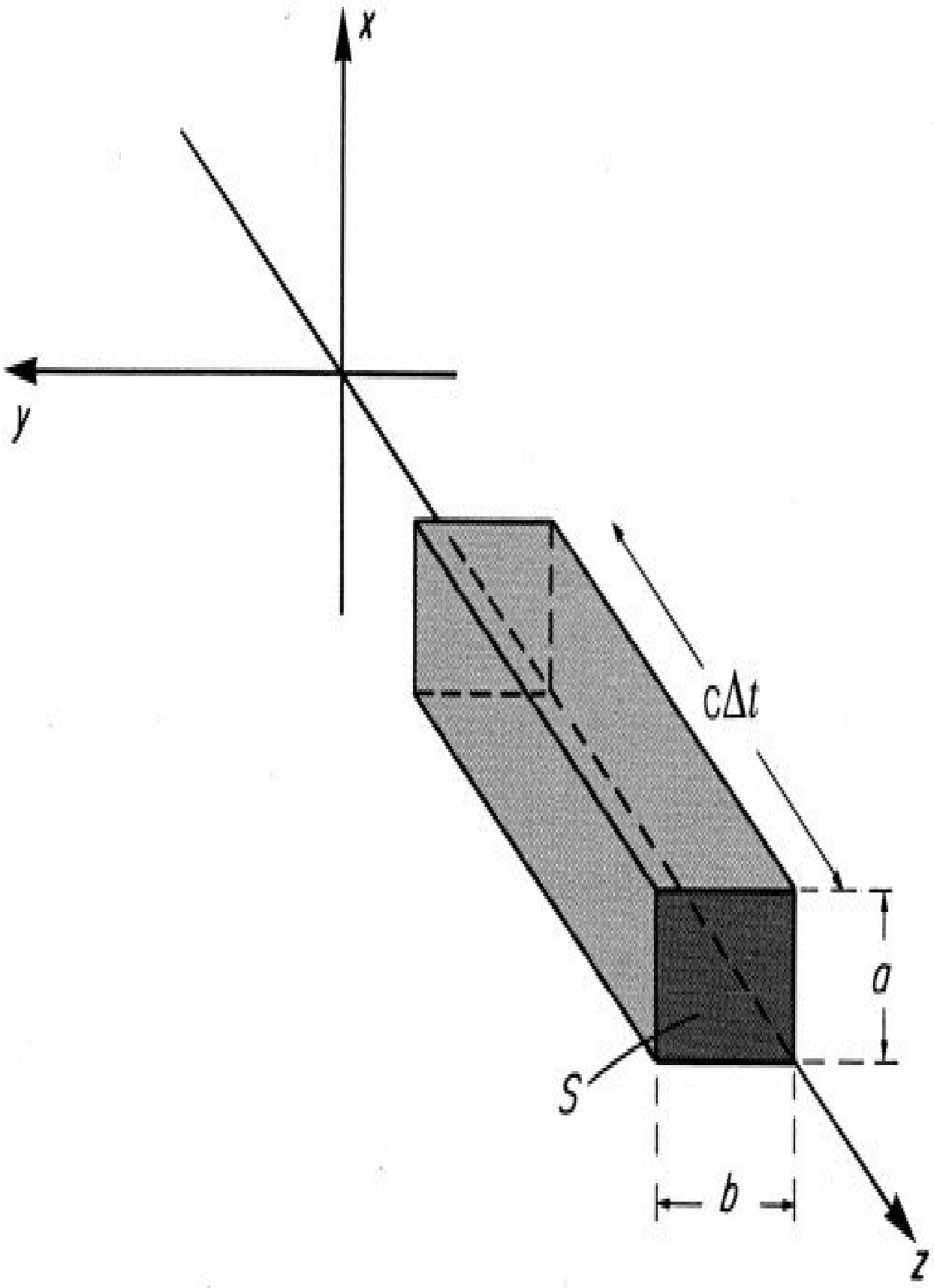
In this section we will obtain an expression giving the energy flowing across unit area per second at any point in an electromagnetic wave. This quantity is a vector which depends on the electric and magnetic fields, and is called the Poynting vector.

There is no dispersion of a wave in free space. The wave velocities of all the different frequency components are the same, and equal to the speed of light in free space, c. The group velocity is also equal to c, and the energy associated with any frequency component in the wave moves in the direction the wave is moving with the same speed. We will calculate the rate of energy flow across unit area associated with a component of angular frequency ω. If the wave is moving in the z-direction the energy flow is in the z-direction and is the same at all points with the same value of the coordinate z. Since the fields are oscillating in time the instantaneous energy flow oscillates also. However the energy flow averaged over a whole period of oscillation of the fields is constant. Figure 11.8 shows a rectangular box lying along the z-axis. The box has sides of lengths a and b along the x- and y-directions, and length in the z-direction given by the product of the speed of light in vacuo and an arbitrary time Δt. The energy contained in the box, averaged over a cycle of the fields, is the average energy crossing the area S in time Δt, since the energy in the wave travels with the speed c.

In sections 2.4 and 6.3.4 we obtained expressions which together give the total energy U in a combination of static electric and magnetic fields.

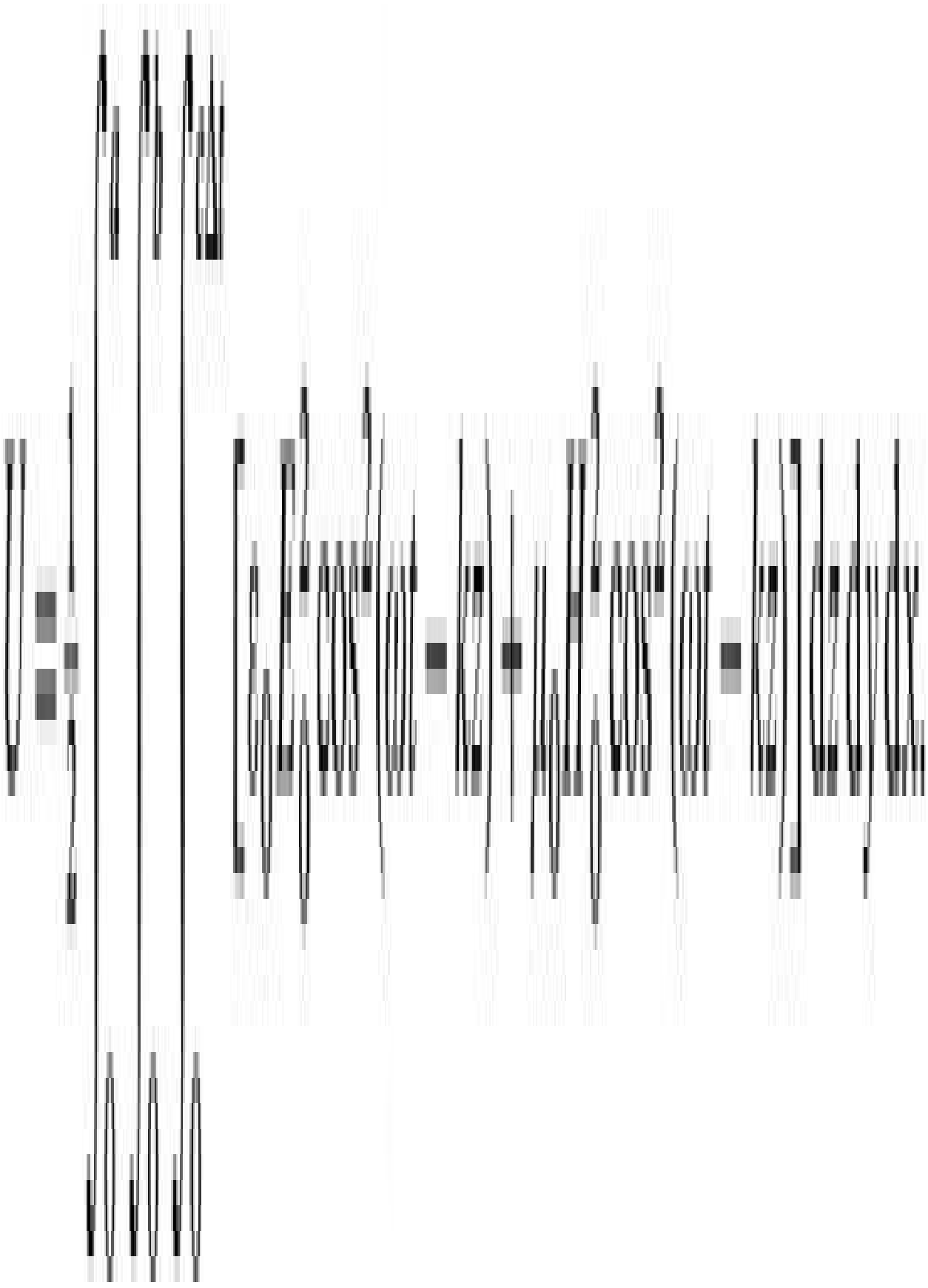
(11.32)

Figure 11.8. When a plane wave propagates in the z-direction in free space, the average energy crossing the surface S in time τ is equal to the average energy stored in the fields within a box of length equal to the speed of light × the time τ.

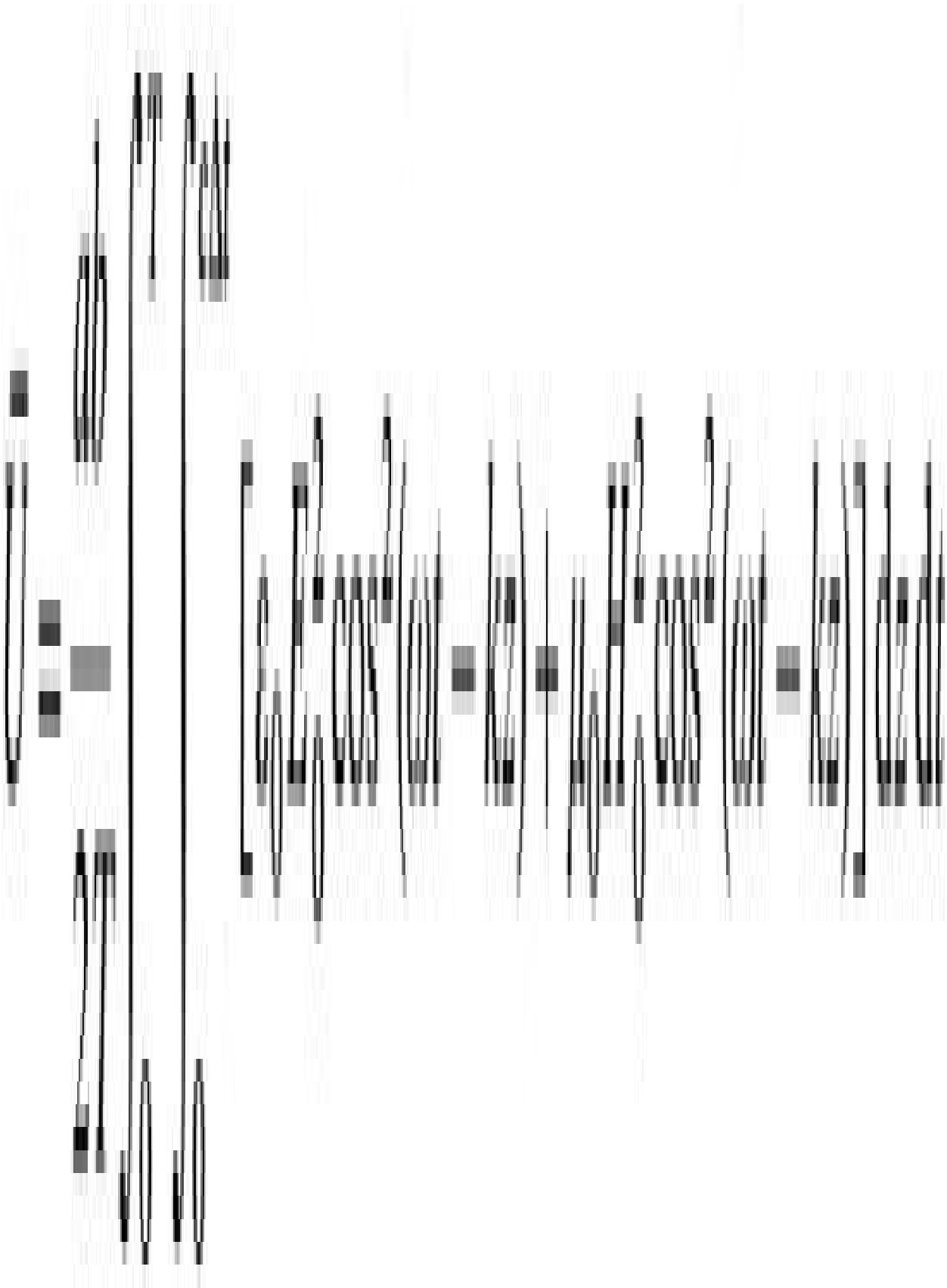

x
y
cΔt
a
S
b
z

where V is the volume over which the fields are non-zero. We will assume that this equation also gives the energy when the fields change with time. Furthermore we will assume that it correctly gives the total energy in an arbitrary volume V. Using Equation (11.32) and the relations $D = \varepsilon_0 E$, $B = \mu_0 H$, we find that at time t the energy density in the box drawn in the figure is equal to $\varepsilon_0 E^2/2 + \mu_0 H^2/2$ where the fields E and H are given by

In evaluating energy, which is quadratic, not linear, in the fields, it is necessary to abandon the complex notation and remember that the fields are given by the real parts of the above expressions. With this change the total energy in the box at time t is given by



The average energy in the box is thus

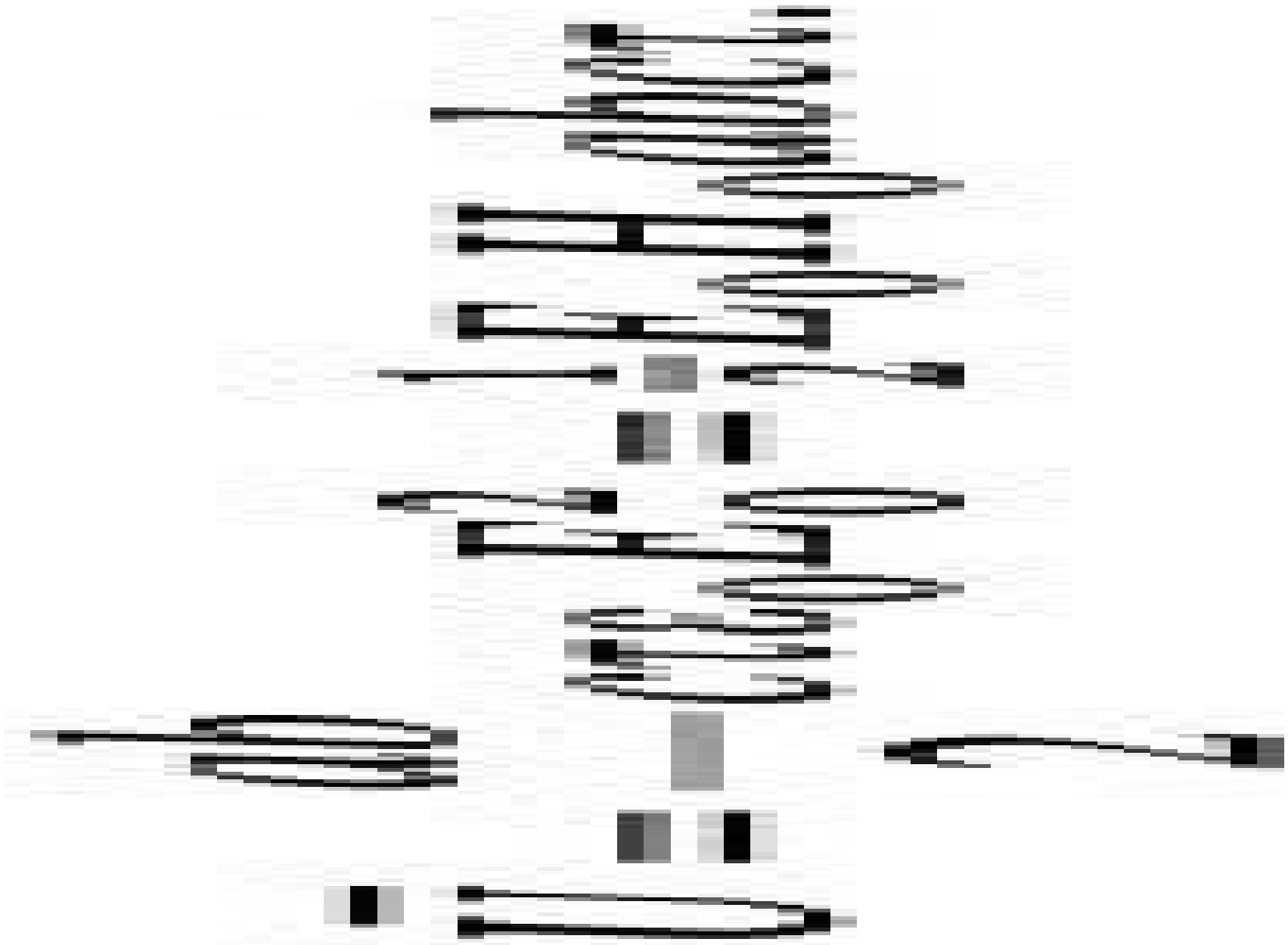


where T is the period of the oscillations and is equal to $2\pi/\omega$. Hence

$$\bar{U} = \frac{ab}{2}\int_0^{c\Delta t}\left(\frac{1}{2}\varepsilon_0 E_0^2 + \frac{1}{2}\mu_0 H_0^2\right)dz$$

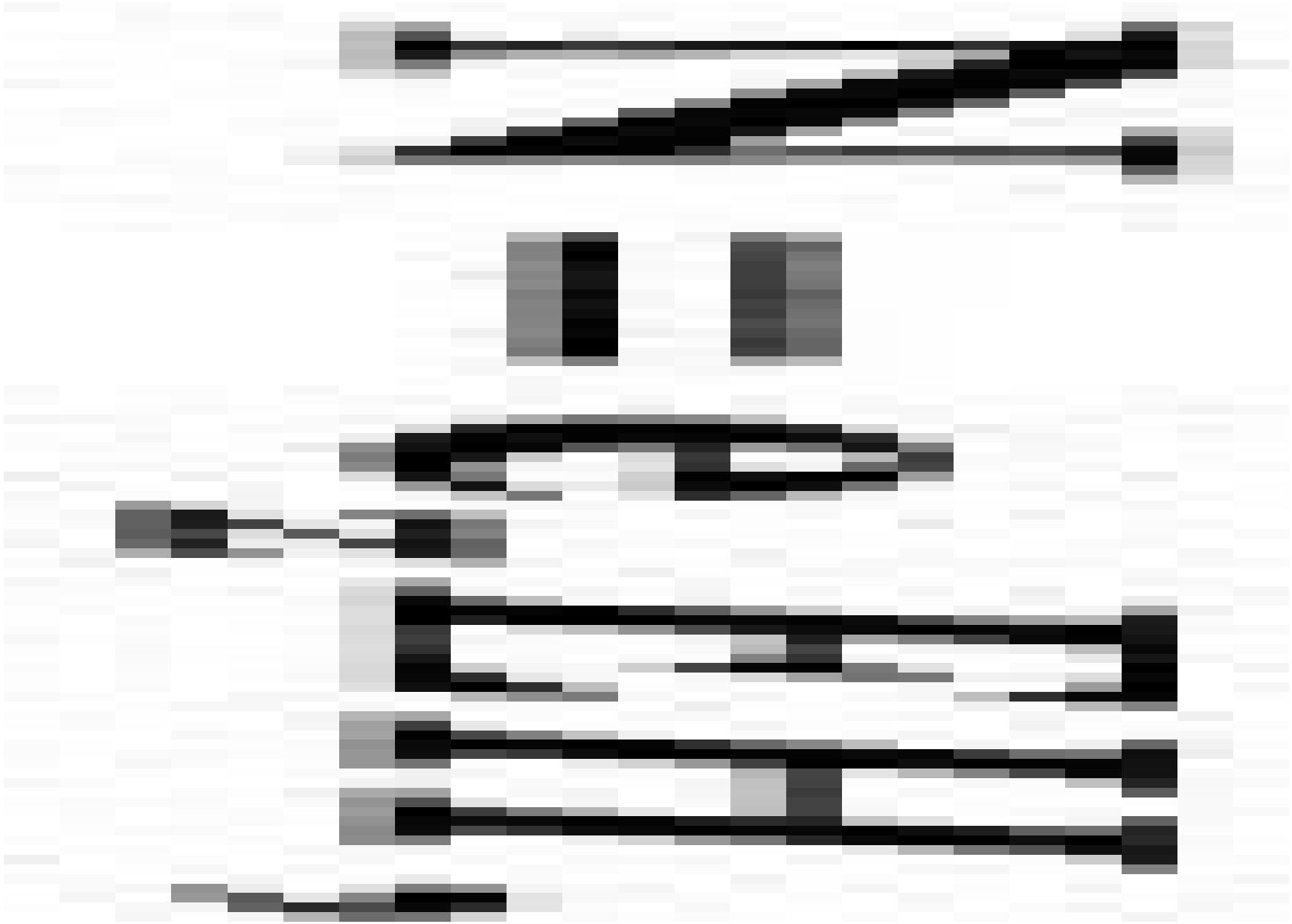
$$= \frac{ab}{4}c\tau\left(\varepsilon_0 E_0^2 + \mu_0 H_0^2\right).$$

Using the relations $H_0 = B_0/\mu_0$, $c = (\varepsilon_0\mu_0)^{-1/2}$ and $B_0 = E_0/c$ from Equation (11.19) the above equation reduces to



The average energy crossing unit area per second at any point in the plane wave is thus E0H0/2, and its direction is the z-direction. Introducing the symbol N for the Poynting vector we thus have

This is the result we would obtain if the instantaneous value of the Poynting vector were



or alternatively

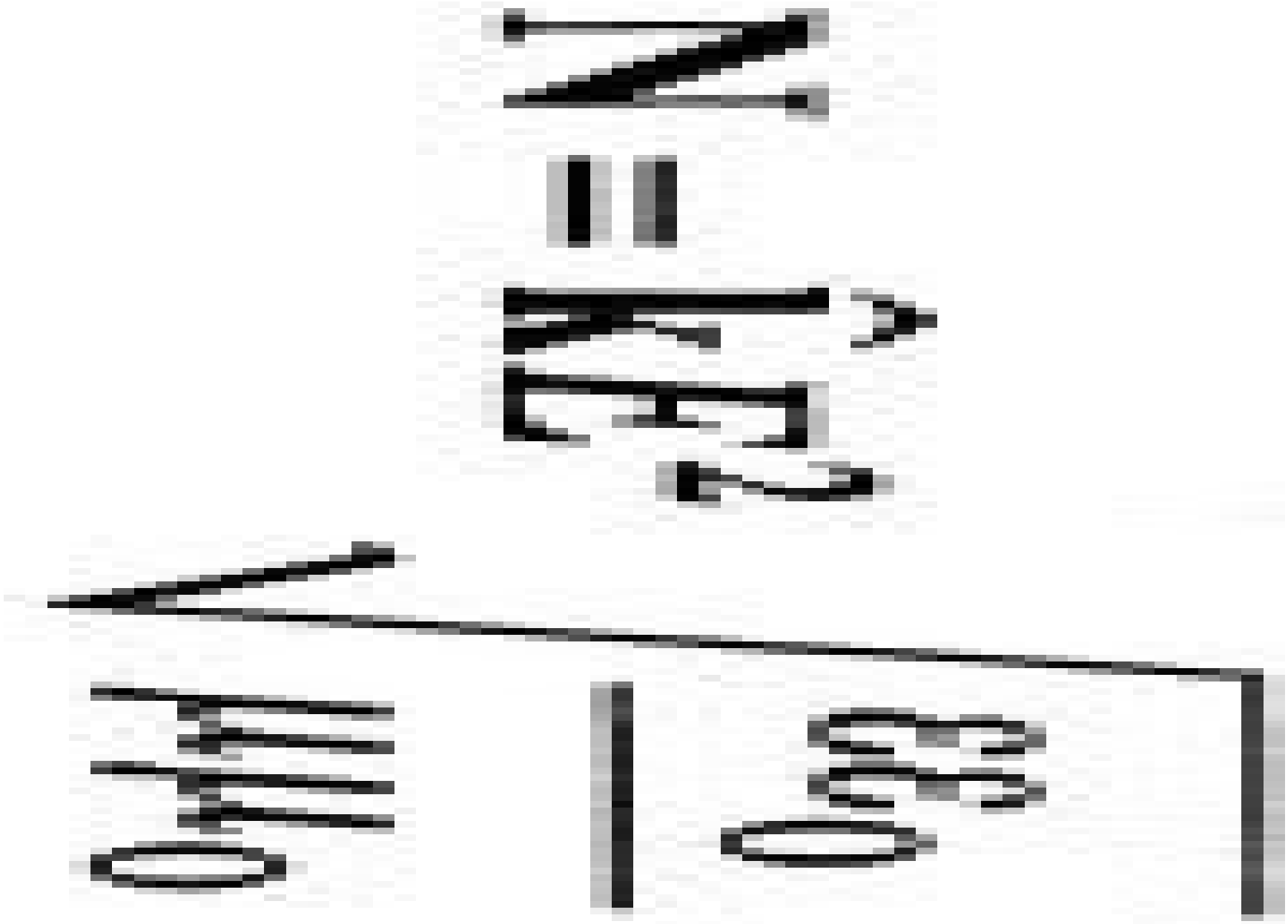
(11.33)

This equation gives the Poynting vector for a plane wave in free space; however it can also be used to obtain the energy flow in any electromagnetic field. It is shown below that the interpretation of expression (11.33) as the energy crossing unit area per second in any electromagnetic field is consistent with the conservation laws of physics and with Maxwell's equations. All experiments so far performed support this interpretation.

For a plane wave moving in the direction of the unit vector in an isotropic medium with the relative permittivity ε and relative permeability μ, the Poynting vector is

$$
\begin{aligned}
\mathbf{N} &= \mathbf{E} \times \mathbf{H} \\
&= \mathbf{E} \times \frac{\mathbf{B}}{\mu\mu_0} \\
&= \mathbf{E} \times (\hat{\mathbf{k}} \times \mathbf{E}) \frac{\sqrt{\varepsilon\varepsilon_0\mu\mu_0}}{\mu\mu_0},
\end{aligned}
$$

from Equations (11.22) and (11.25). This gives



or

(11.34)

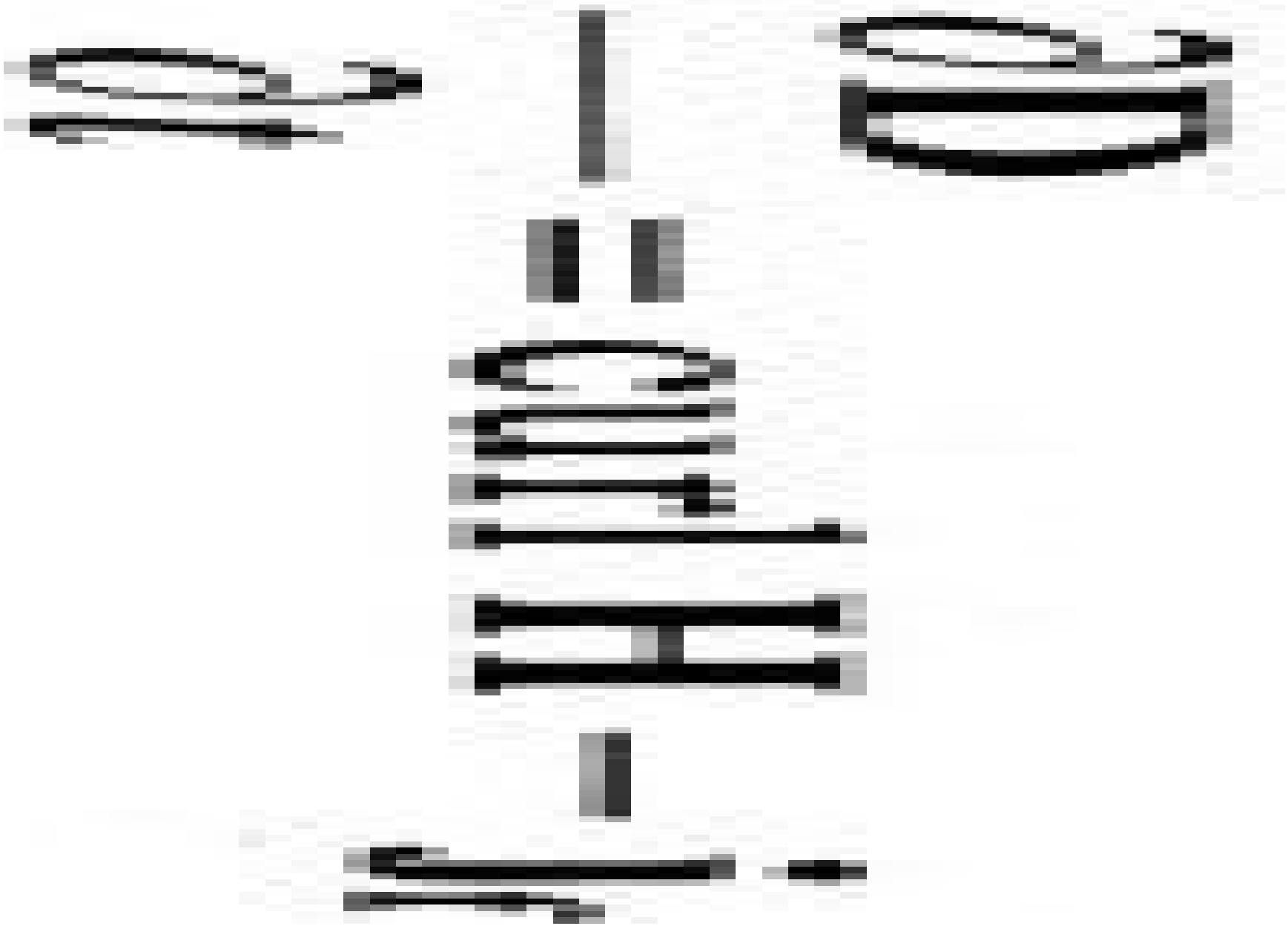
If the amplitude of the electric field in a plane wave is 0.1 V m−1 (a typical value a few km from a radio transmitter) the energy crossing unit area per second in air is l0−2 . If the radio wave were being propagated in water (which has a relative permittivity of about 80 at radio frequencies), and at some point the electric field had the amplitude 0.1 V m−1, the Poynting vector would be increased to ~ 1.2 × 10−4 W m−2.

We now outline a more general, but more mathematical approach to the Poynting vector in linear isotropic materials. The remainder of this section can be omitted without loss of continuity. We begin with Equation (11.32) which gives the energy in a volume V enclosed by a surface S. The material inside the surface S is taken to be uniform and isotropic and to have relative permittivity ε and relative permeability μ. The rate of change of energy inside the surface S is equal to

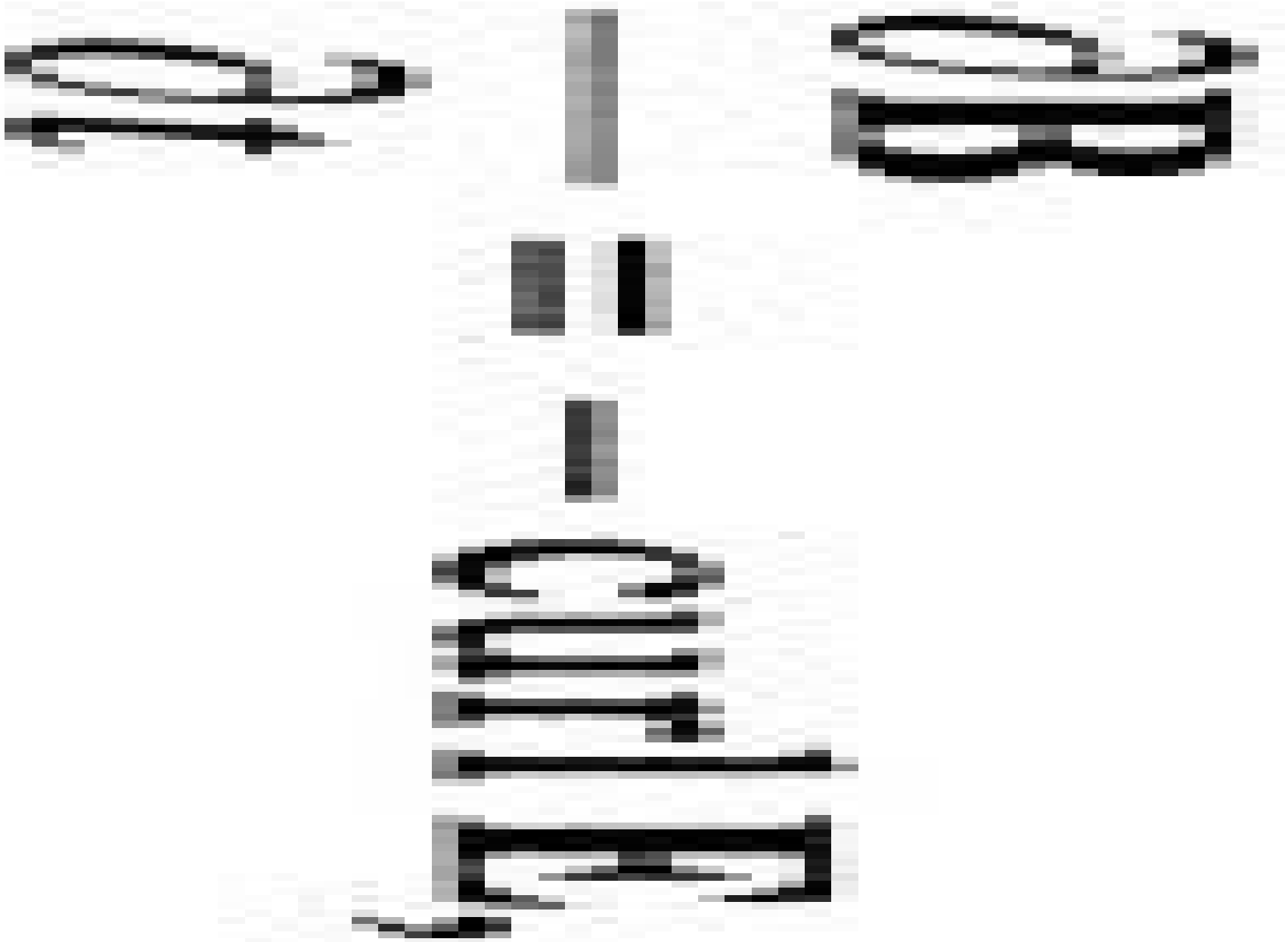
$$\frac{\partial U}{\partial t} = \frac{1}{2}\frac{\partial}{\partial t}\left(\int_V (\mathbf{E}\cdot\mathbf{D} + \mathbf{B}\cdot\mathbf{H})\,d\tau\right)$$

$$= \int_V \left(\frac{\partial \mathbf{D}}{\partial t}\cdot\mathbf{E} + \frac{\partial \mathbf{B}}{\partial t}\cdot\mathbf{H}\right) d\tau.$$

But

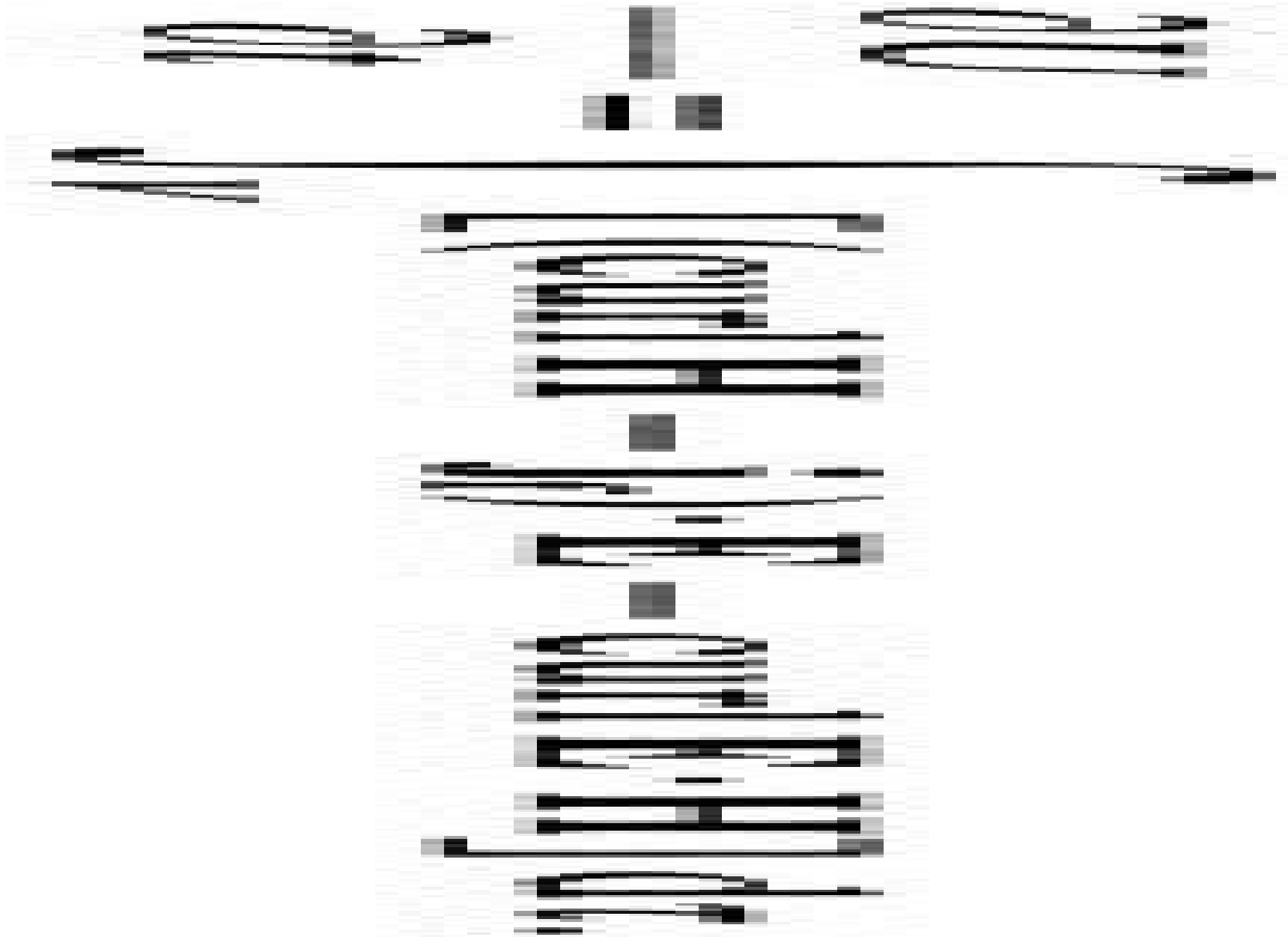


and

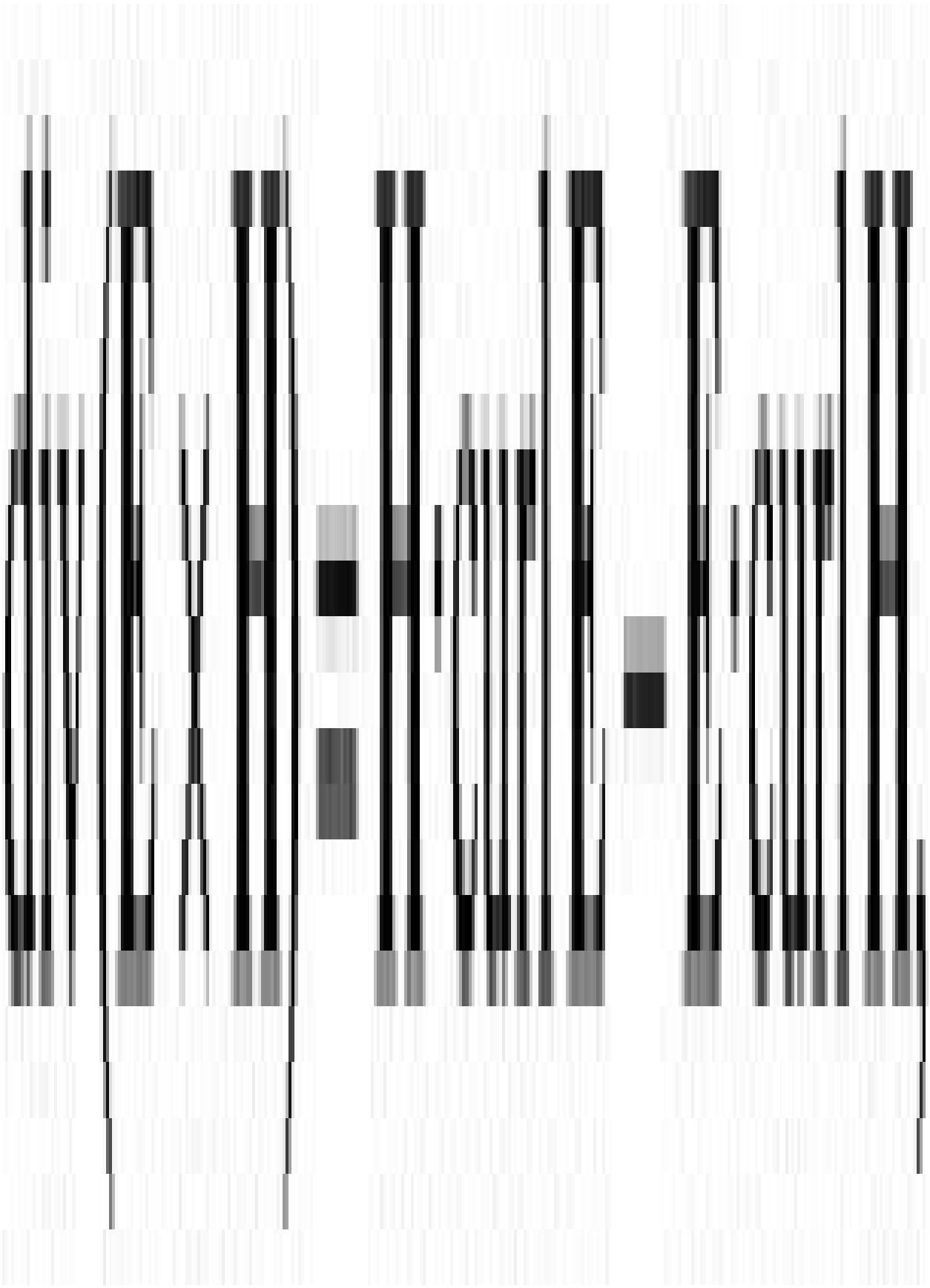


from Maxwell’s equations (10.13) and (10.12). Hence

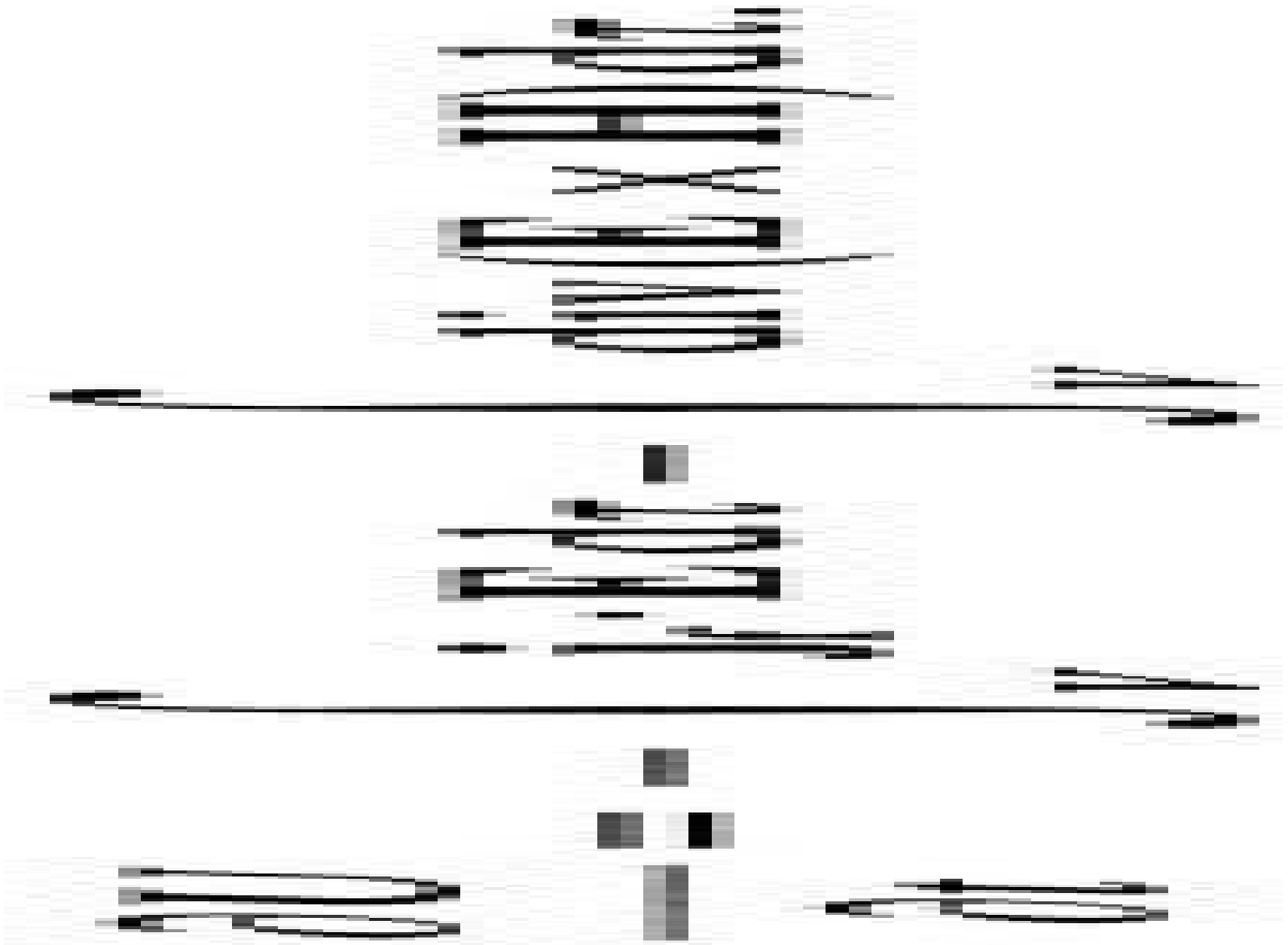
(11.35)



Using the vector identity (B.24),

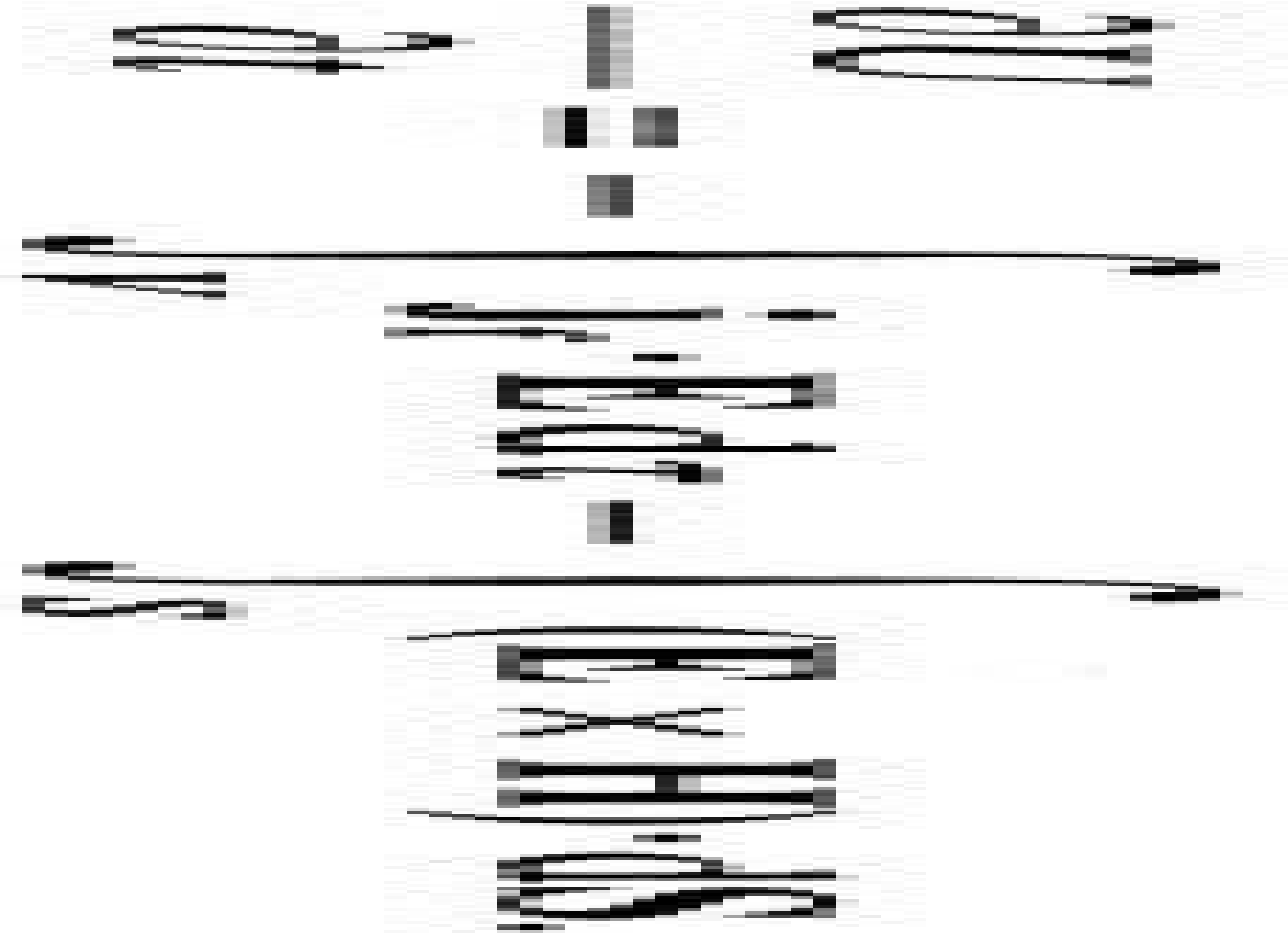


Equation (11.35) can be rewritten as

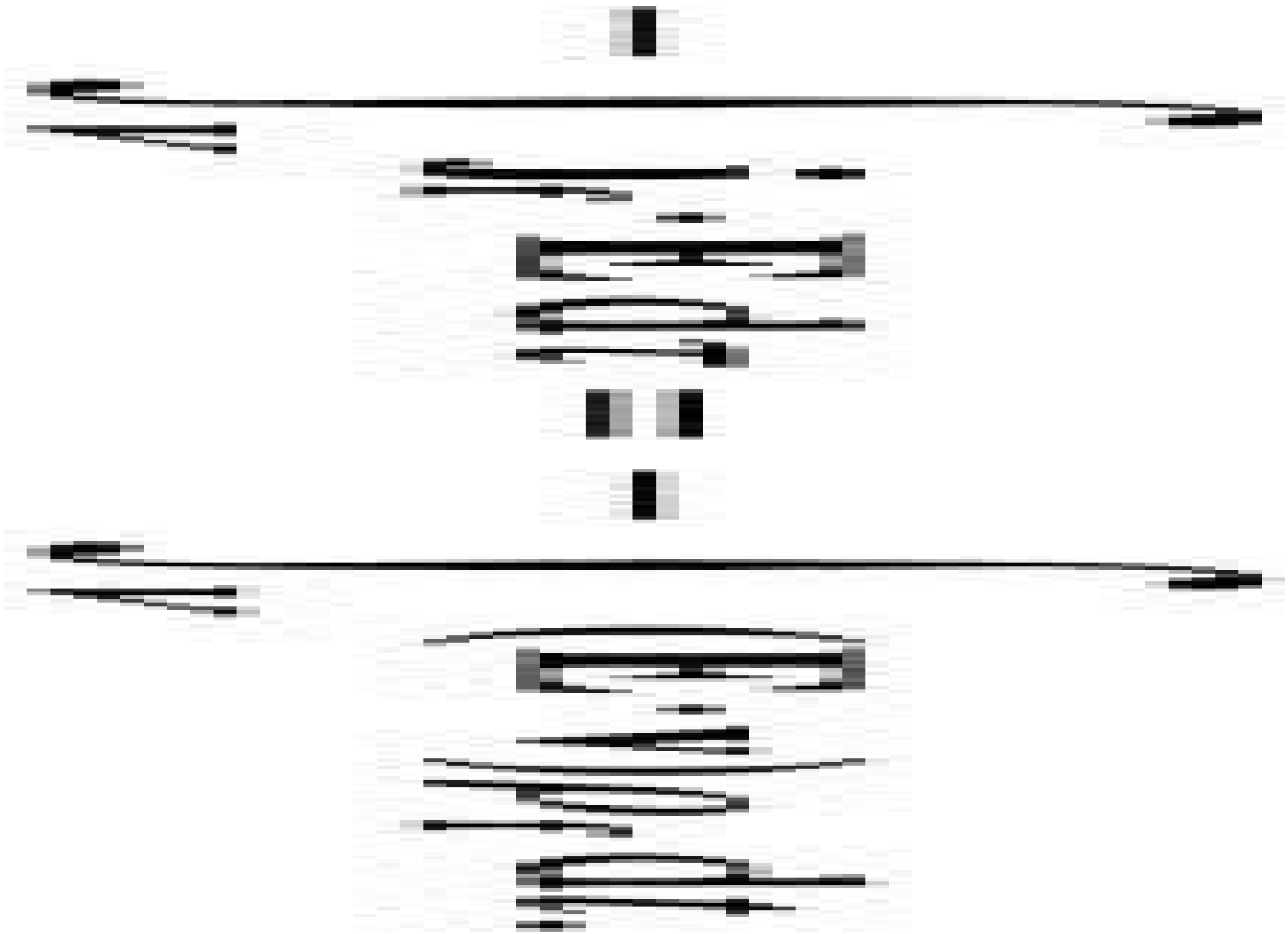


The last term on the right-hand side of this equation can be transformed into an integral over the surface S by using the divergence theorem, and we have

(11.36)



If the free charge density within the volume V is given by ρf, and if the average velocity of the charges is v, the current density jf is equal to ρfv, and hence



The term on the right-hand side of this equation is equal to the rate at which the electric field E does work on the free charges within the volume V. This work appears as heat, and the first term on the right-hand side of Equation (11.36) thus represents the rate of Joule heating in the volume V. The energy in the electromagnetic field in the volume V decreases at a rate equal to the rate of Joule heating plus the energy flowing out of the surface S per unit time. Hence we are led to interpret the integral $\int_S (E \times H) \cdot dS$ as the amount of energy flowing outwards over the surface S per second. This interpretation is consistent with the conservation of energy and the original assumption that the energy was given by Equation (11.32), i.e. has the same form as for static fields. We now go one step further and postulate that the vector $N = E \times H$ represents the energy flow per unit area per second at every point.

11.5. THE ABSORPTION OF PLANE WAVES IN CONDUCTORS AND THE SKIN EFFECT

If a wave is propagating in a conducting medium the oscillating electric field in the wave sets up currents. Work must be done to drive the currents, and some of the energy in the wave is dissipated as heat in the medium. A plane wave progressing through a conducting medium must therefore be attenuated. In this section we determine the attenuation and discuss some of its practical consequences.

We will consider materials for which the current density j is proportional to the electric field E, and write

(11.37)

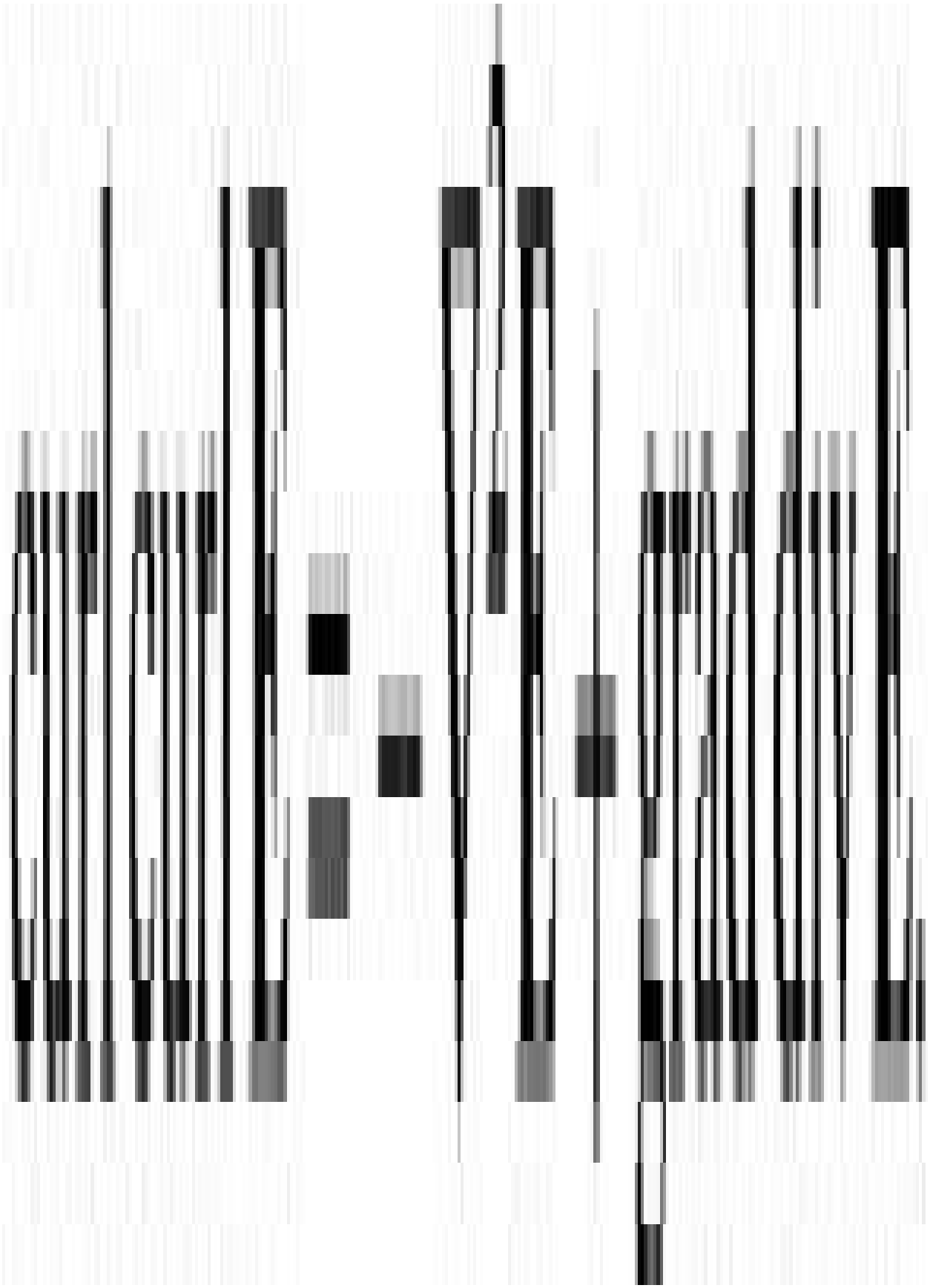
where the constant σ can be taken to be equal to the steady current electrical conductivity. This relationship is almost exact, up to the highest frequencies where classical theory is applicable, for most materials that obey Ohm's law with direct currents. For very rapidly changing currents, the relationship may no longer be an adequate approximation (see Problem 10.2). We now use Maxwell's equations to derive an expression which represents an electromagnetic wave propagating in a conducting medium in which the relationship (11.37) is obeyed. In this case the last two Maxwell Equations (10.12) and (10.13) can be written

(11.38)

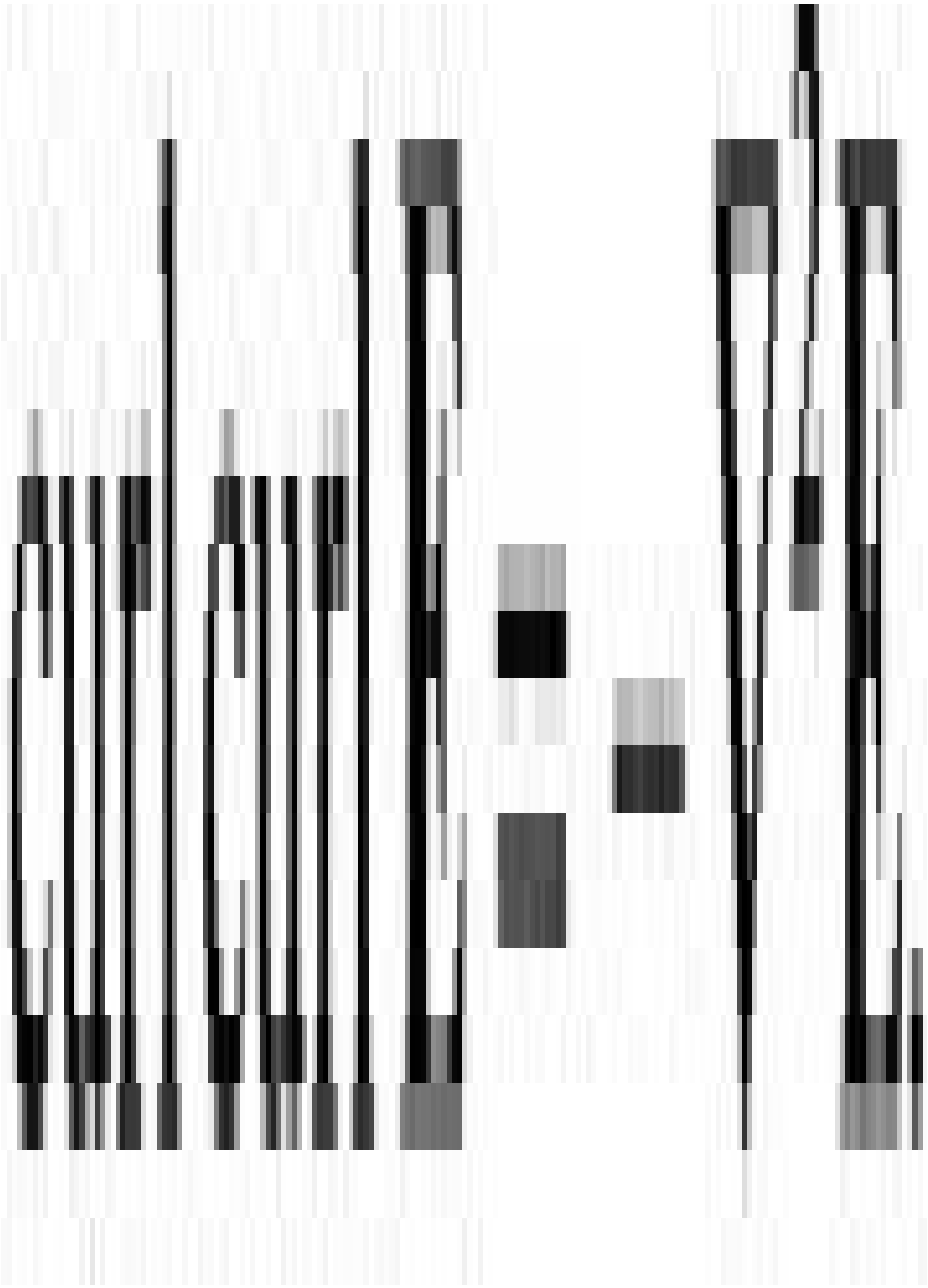
and

(11.39)

where ε and μ are the relative permittivity and permeability respectively. From these two equations we can derive an equation involving the field E alone, in a similar manner to the procedure used in section 11.1 for waves in free space. We again use the relation (B.21),



When a wave is travelling through a conducting medium the macroscopic charge density remains zero everywhere, and so we have div E = 0. The last equation under these conditions becomes



From Equation (11.38)

$$\operatorname{curl}\operatorname{curl}\mathbf{E} = -\operatorname{curl}\left(\frac{\partial \mathbf{B}}{\partial t}\right)$$

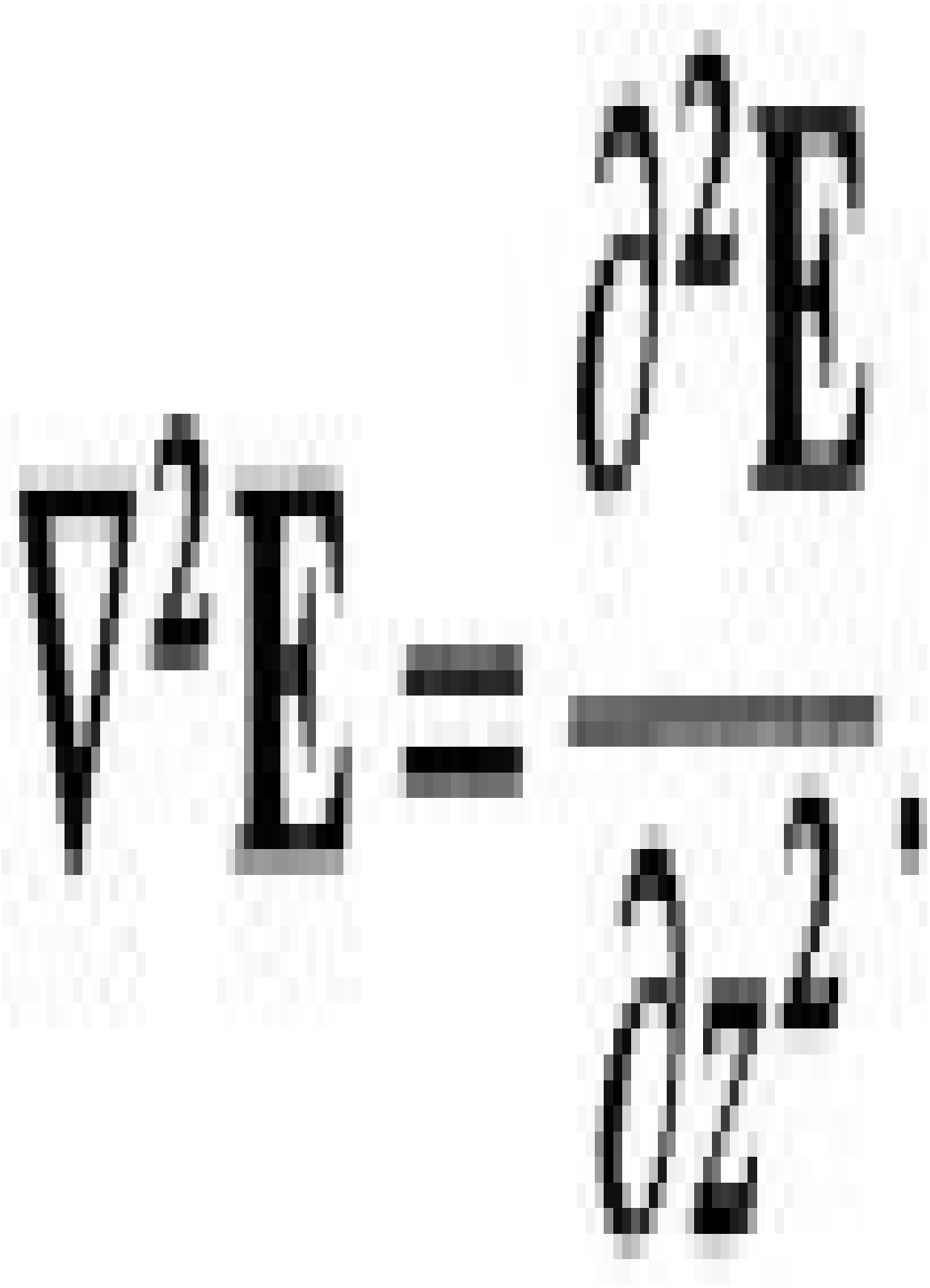
$$= -\frac{\partial}{\partial t}(\operatorname{curl}\mathbf{B}),$$

and hence

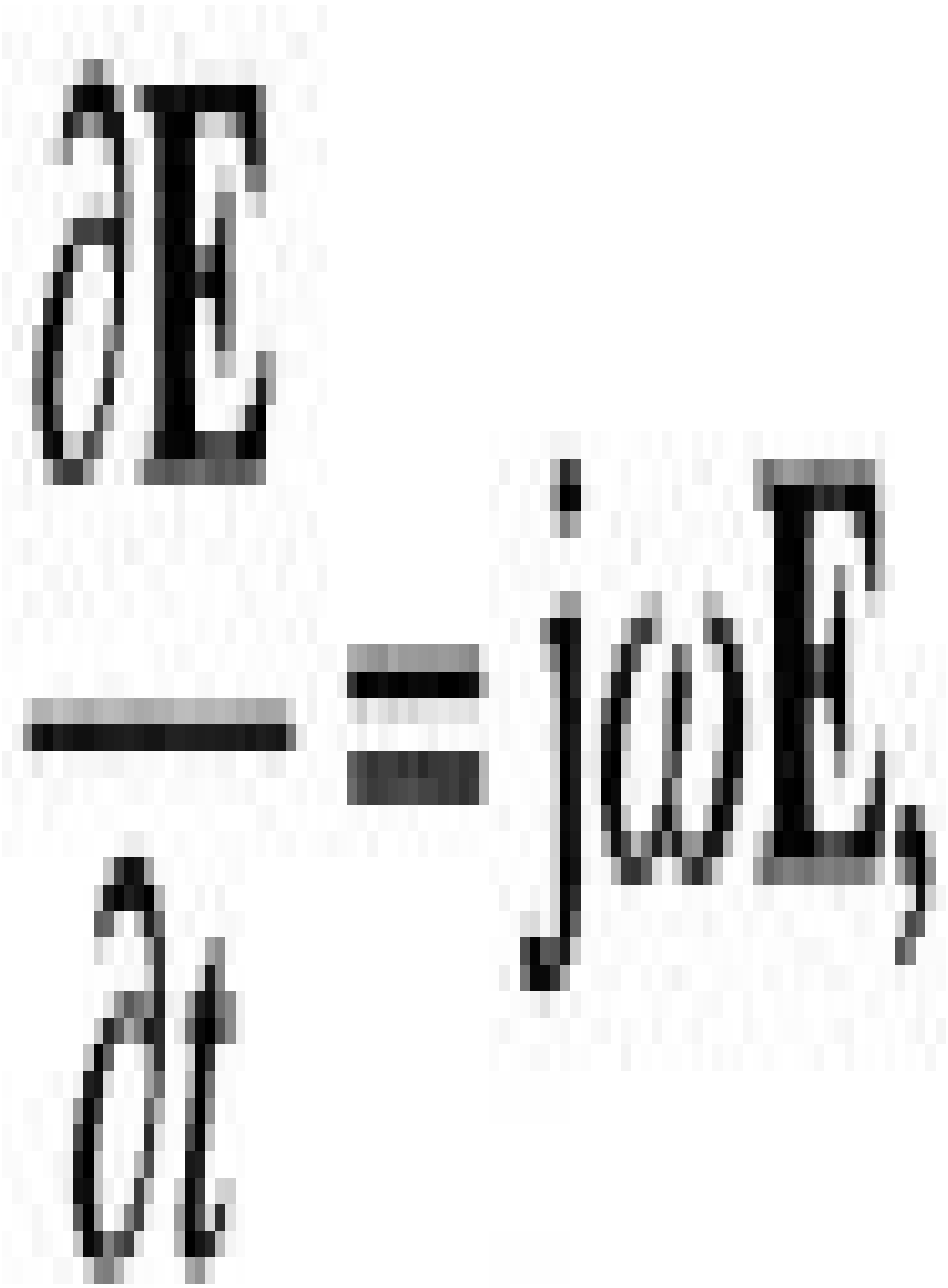
(11.40)

from Equation (11.39).

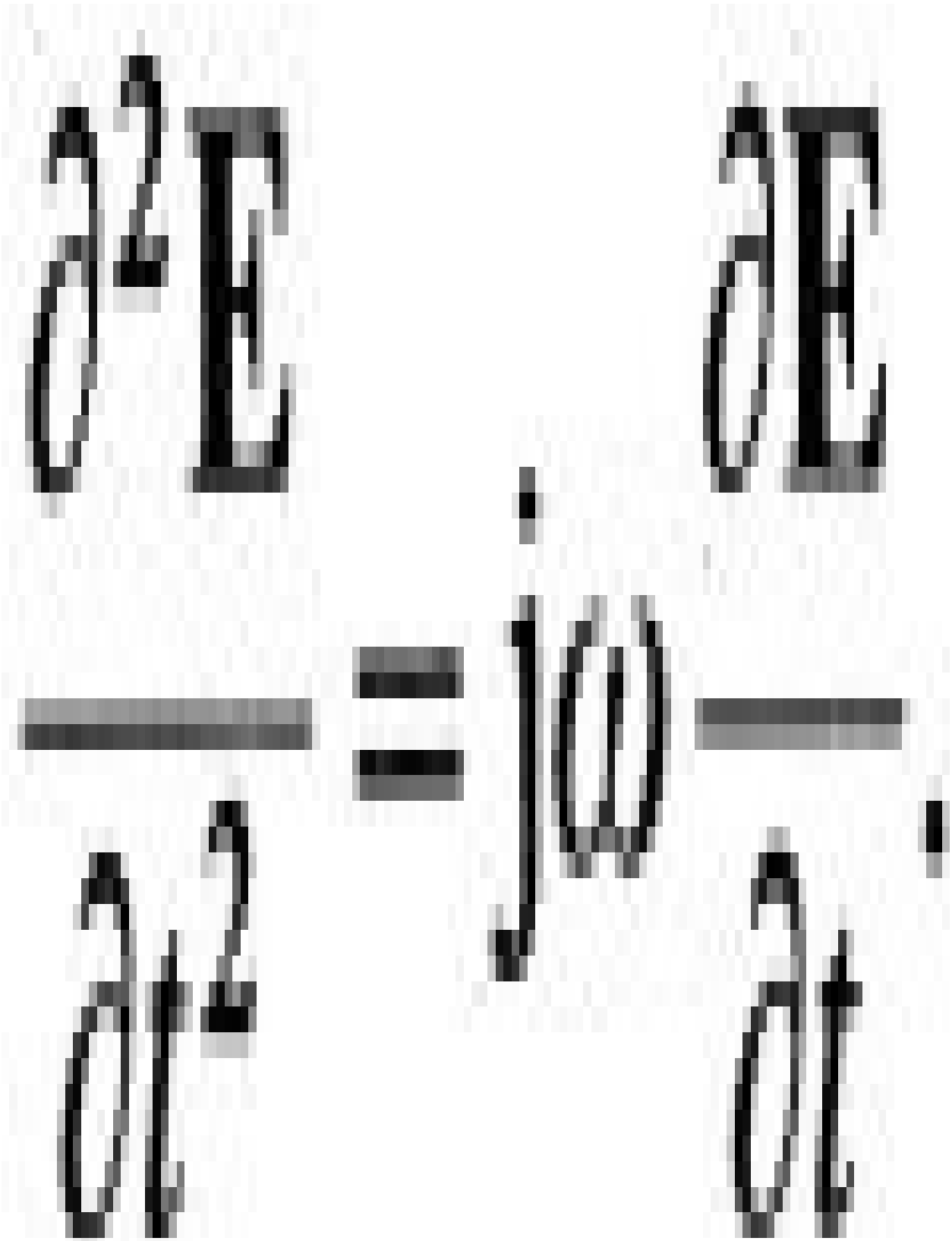
Let us look for a solution to Equation (11.40) which represents an attenuated plane wave moving in the z-direction, i.e. one in which the fields are constant everywhere in the x–y plane at a fixed value of the coordinate z, but whose amplitudes decrease exponentially as z increases. In such a wave



If the wave has angular frequency ω,



and



Substituting these expressions into Equation (11.40) gives the equation

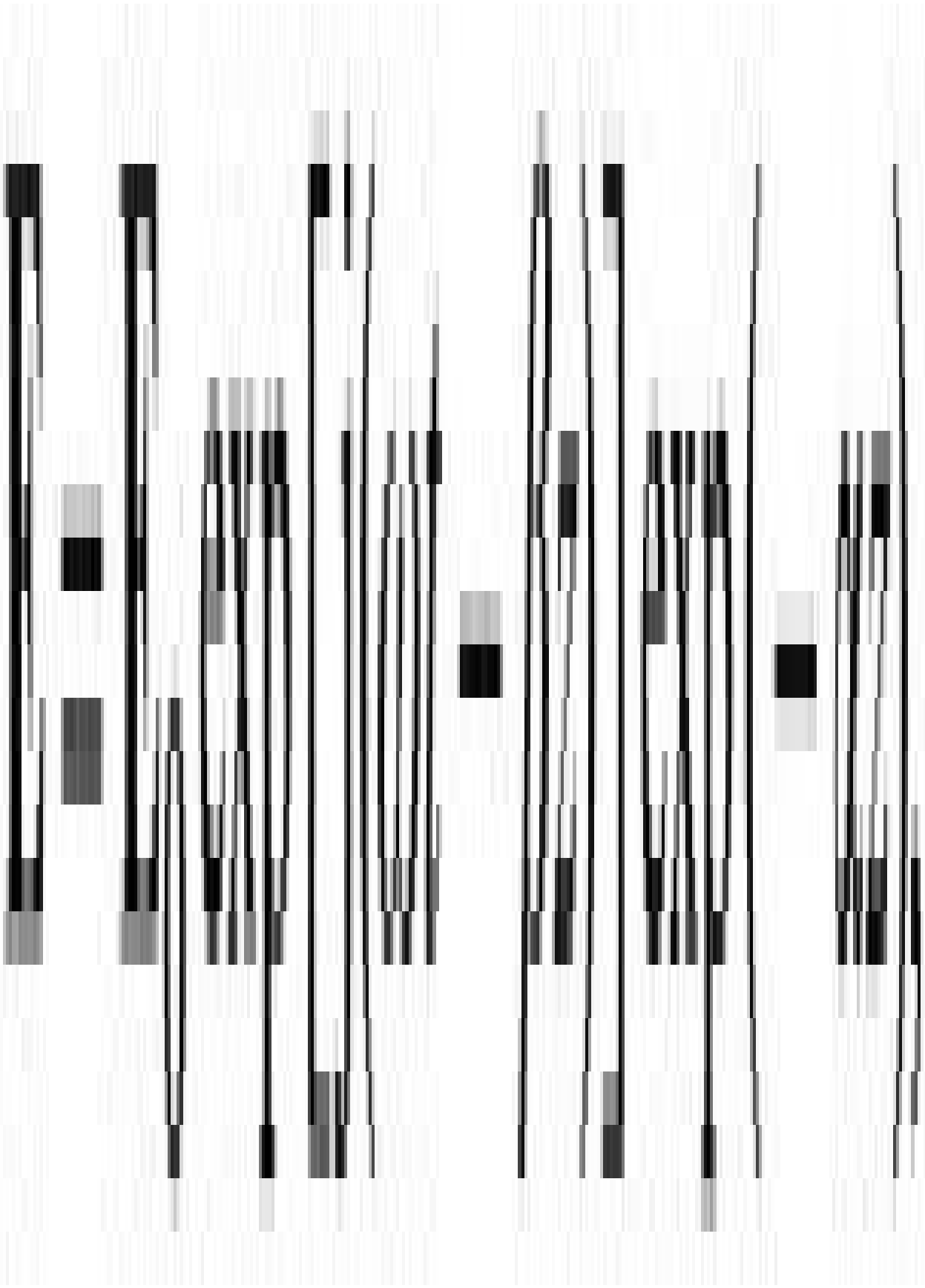
(11.41)

The first term on the right-hand side arises from the displacement current; the second term from the current due to motions of real charges. For almost all conducting media at the frequencies of interest the product $\omega\varepsilon\varepsilon_0$ is very much less than the electrical conductivity σ. For example, a weak electrolyte would have a conductivity of about 1 $(\Omega\ \mathrm{m})^{-1}$. At a frequency of 1 MHz the value of $\omega\varepsilon\varepsilon_0$ is less than 10^{-2} $(\Omega\ \mathrm{m})^{-1}$. Thus it is usually a good approximation to neglect the displacement current effects on the right-hand side of Equation (11.41). The differential equation obeyed by a plane wave in conducting media then becomes

(11.42)

The attenuated plane wave solution we are looking for has the form

(11.43)



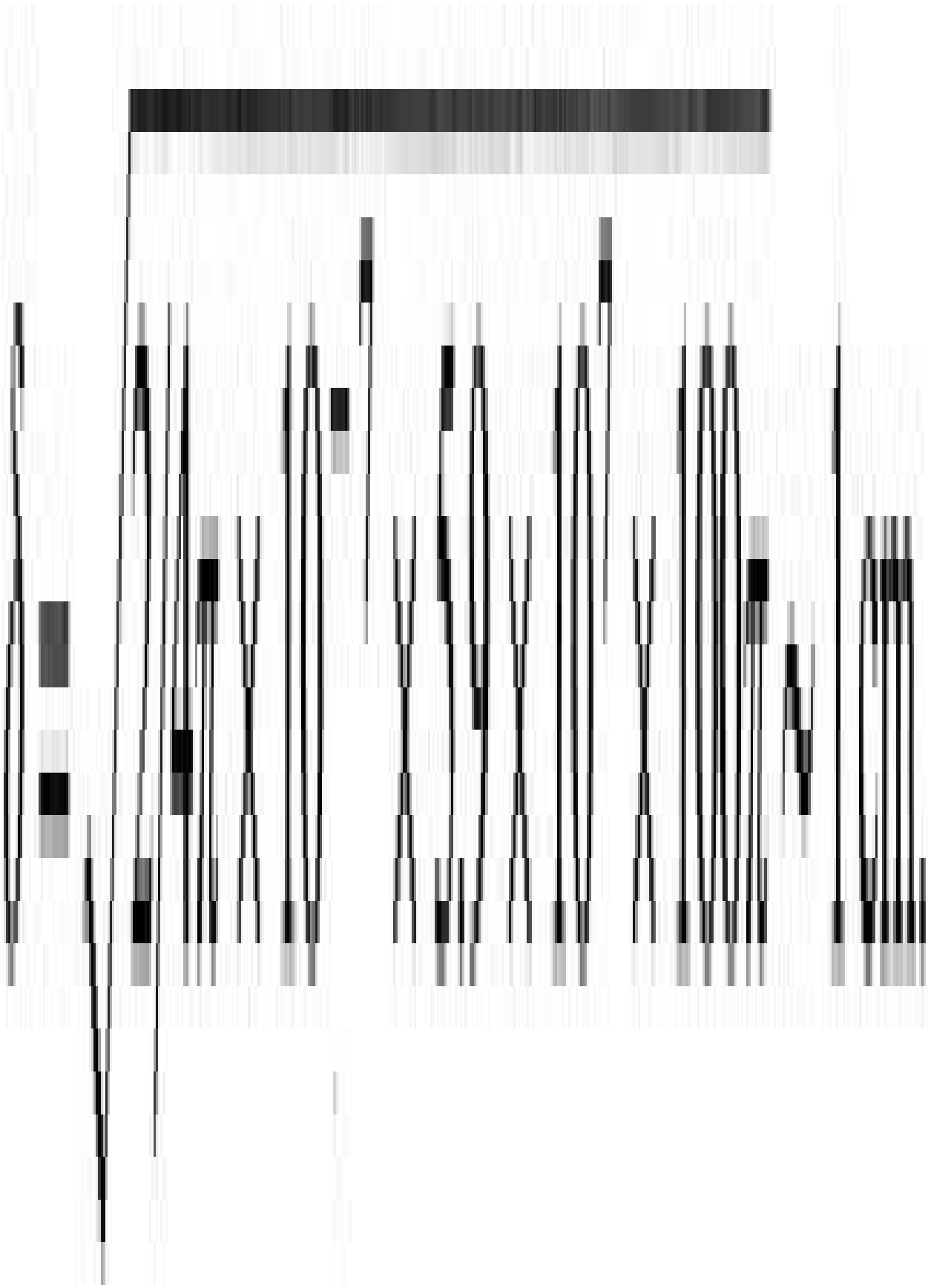
where α and β are parameters to be determined. Substitution of the derivatives obtained from this equation into Equation (11.42) gives

(11.44)

The quantity

(11.45)

measures how rapidly the wave is attenuated and is called the skin depth. Usually the relative permeability μ is very nearly unity and . For a frequency of 50 Hz, the skin depth in copper is



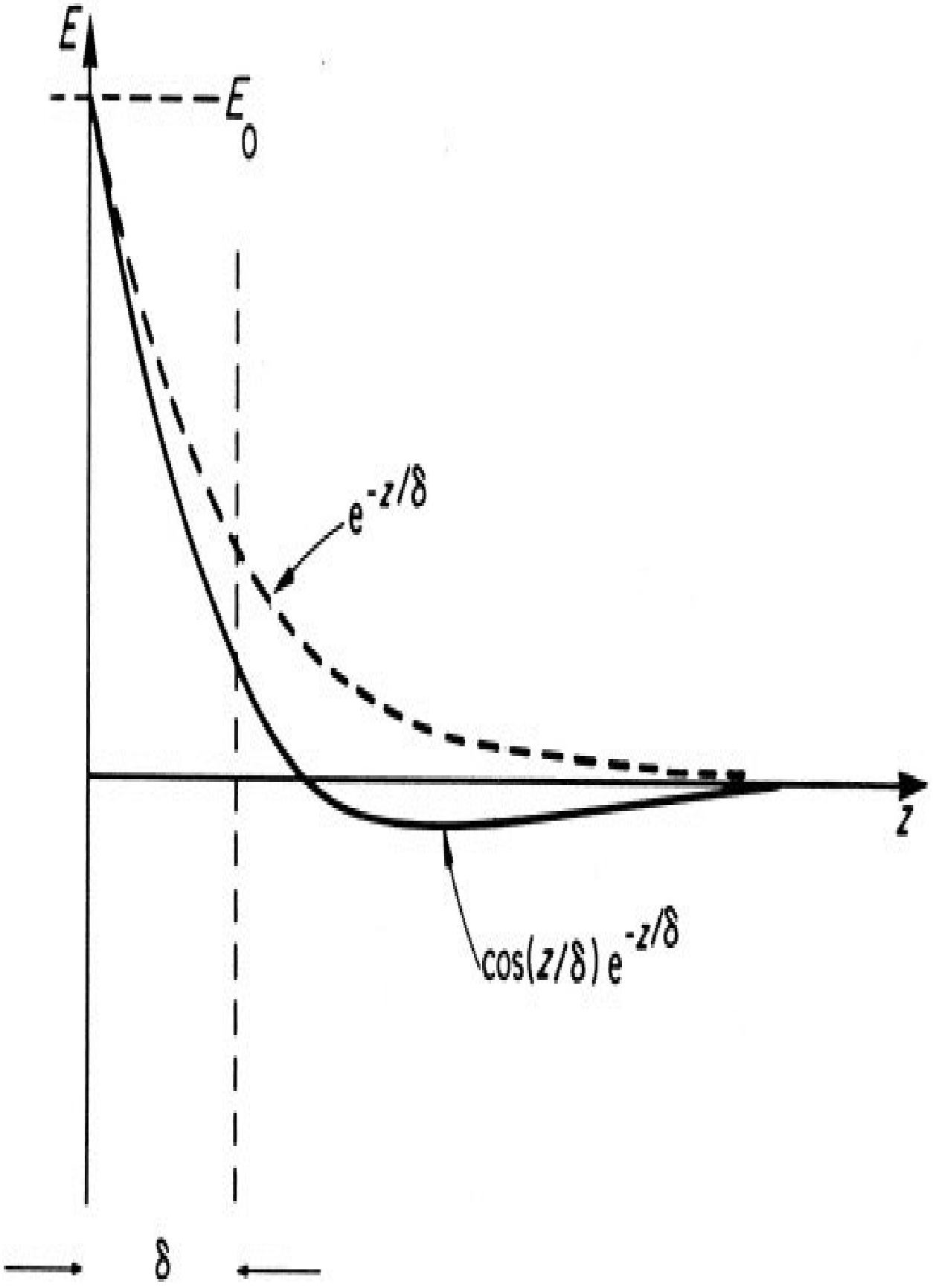
As the frequency increases the skin depth decreases proportionally to $\omega^{-1/2}$ and at a frequency of 50 MHz the skin depth in copper has decreased to $\sim 10^{-3}$ cm.

The fact that the skin depth is small for good conductors at high frequencies has important consequences for the design of conductors to take high frequency currents. Suppose that an alternating current is passed along a long cylindrical conductor. The electric field is everywhere parallel to the axis of the cylinder, and by symmetry depends only on the distance z perpendicularly inwards from the surface. The exact determination of the fields requires the solution of Maxwell's equations fitted to the appropriate boundary conditions. However when examining the behaviour of the electric field near the surface of the conductor, it is a good approximation to consider the surface flat if the skin depth is small compared to the radius of the cylindrical wire. The field will be given in this approximation by Equations (11.43) and (11.44), and the amplitude of the current density $j = \sigma E$ in the rod decreases exponentially with distance inwards as shown in Figure 11.9.

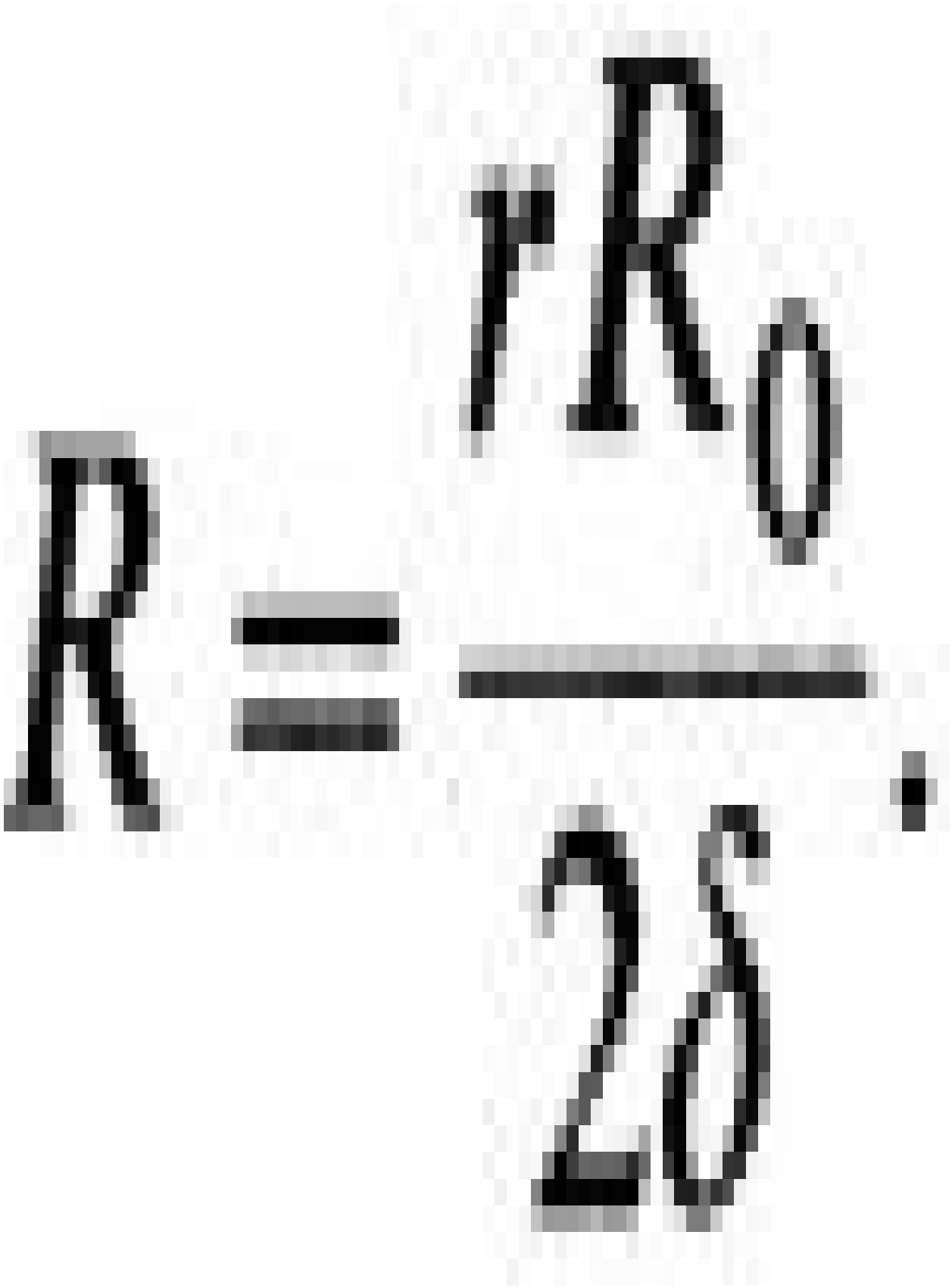
At very high frequencies the current is all carried in a thin outer layer of any conductor. This phenomenon is known as the skin effect. At high frequencies the resistance R of a conductor whose dimensions are much greater than the skin depth δ can be estimated by assuming that all the current is carried uniformly in a surface layer of thickness δ. If the conductor is a cylinder of length l and radius r, its resistance in this approximation is

(11.46)

Figure 11.9. The variation of the electric field with distance inwards from the surface of a cylindrical conductor. The full curve is expression (11.43) for the field at time t = 0. The dotted curve is the locus of the electric field amplitude, i.e. a plot of the function $E_0 \exp(-z/\delta)$


E
E0
e-z/δ
z
cos(z/δ) e-z/δ
δ

If R0 is the resistance of the rod to direct current, then



The resistance to direct current of a copper rod 5 mm diameter is about $8 \times 10{-}4\Omega$ per metre. At a frequency of 50 MHz it is about 0.1 Ω per metre. It would be an expensive waste of copper to use such a conductor at 50 MHz.

11.6. THE REFLECTION AND TRANSMISSION OF ELECTROMAGNETIC WAVES

When plane waves are incident on a boundary between different media, some of the incident energy crosses the boundary and some is reflected. In this section we will determine the reflection and transmission coefficients in a few examples of practical importance. These coefficients depend on the relative permittivities and permeabilities of the media on either side of the boundary, i.e. on their refractive indices.

When static fields were being discussed we found that they had to obey certain conditions at the boundaries between different media. The solutions to the time independent Maxwell equations within regions of different permittivities or permeabilities had to be fitted to each other at the boundaries between the different regions, and this fitting procedure enables the fields everywhere to be uniquely determined. The same procedure has to be followed when the fields change with time, and we will have to digress a little in order to discuss the boundary conditions that the changing fields have to obey.

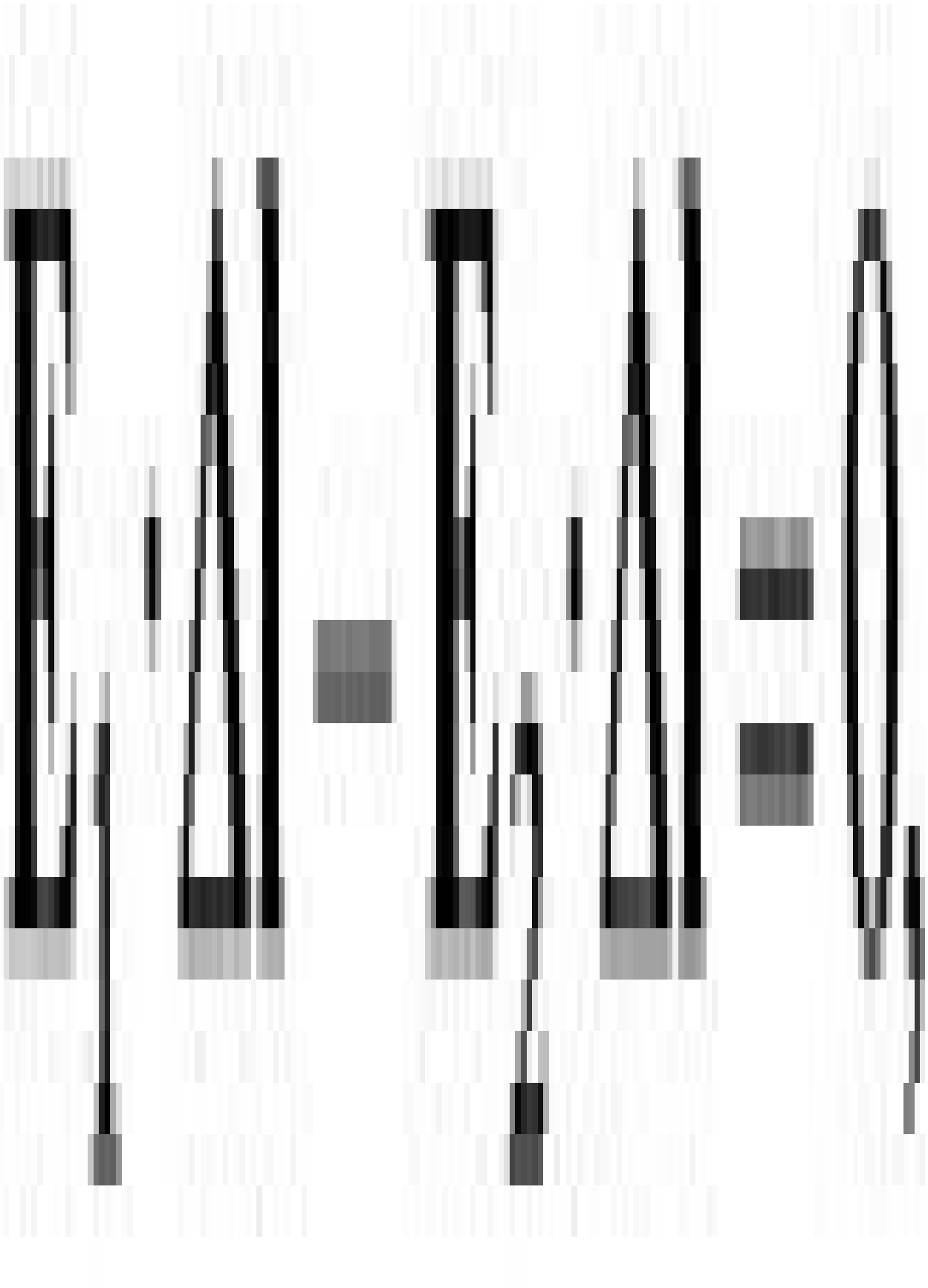
11.6.1 Boundary conditions on electric and magnetic fields

(i) The tangential component of the electric field E

The integral form of Faraday's law is

(11.47)

where S is a complete surface spanning the loop over which the line integral is taken, and the direction of the vector dS is given by the right hand rule. Let us apply Equation (11.47) to the small surface S shown in Figure 11.10 which cuts the interface between the two different media and is perpendicular to it. The sides AD and BC cutting the surface are of equal length Δt. The sides AB and DC, parallel to the interface just above and below it, are of equal length Δl. The vectors E1 and E2 lie in the same plane as the rectangle ABCD. Let us make the length Δt smaller and smaller, i.e. tend to zero. The area of the surface S then tends to zero. The magnetic field B and its derivatives are always finite in real situations involving finite charges and currents, so that the integral on the right-hand side of Equation (11.47) tends to zero also. Hence



where Δl is a vector in the direction from A to B, and E1 and E2 are the fields along AB just above the interface, and along DC just below it. Hence the tangential components of the electric field are continuous at the boundary, and the condition

(11.48)

which we found earlier to be true for the electrostatic field, is true for the electric field in general.

(ii) The tangential component of the magnetic intensity H

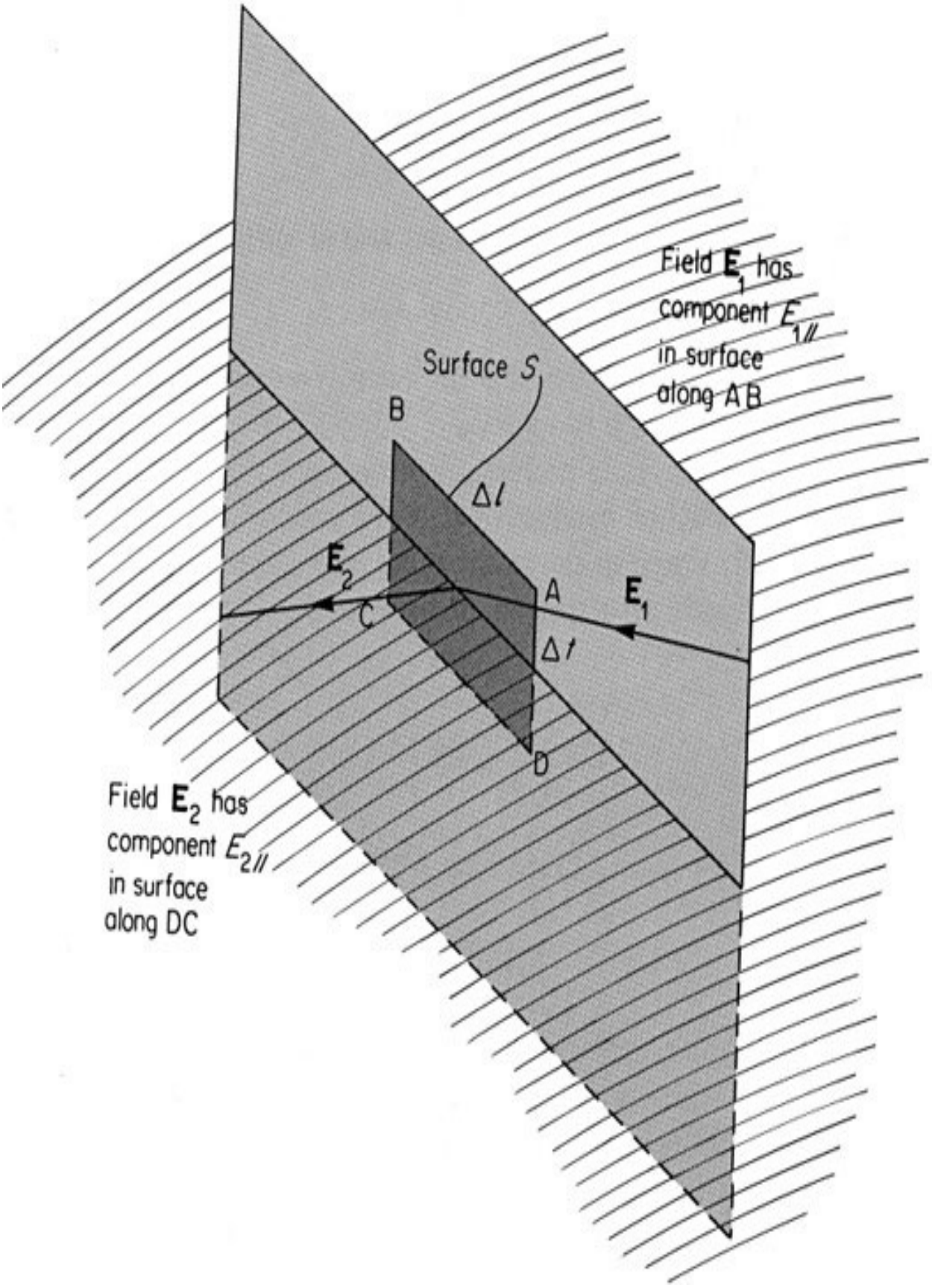
The integral form of the Maxwell equation relating the fields H and D is

(11.49)

Let us continue to use Figure 11.10 where the sides AB and DC are now parallel to the projections in the surface of the fields H1 and H2 on either side, i.e. the vectors H1 and H2 lie in the same plane as the rectangle ABCD. Let the area S shrink as before. The second term on the right-hand side of Equation (11.49) tends to zero as the length Δt tends to zero, since the field D and its derivatives are everywhere finite. The first term on the right-hand side becomes

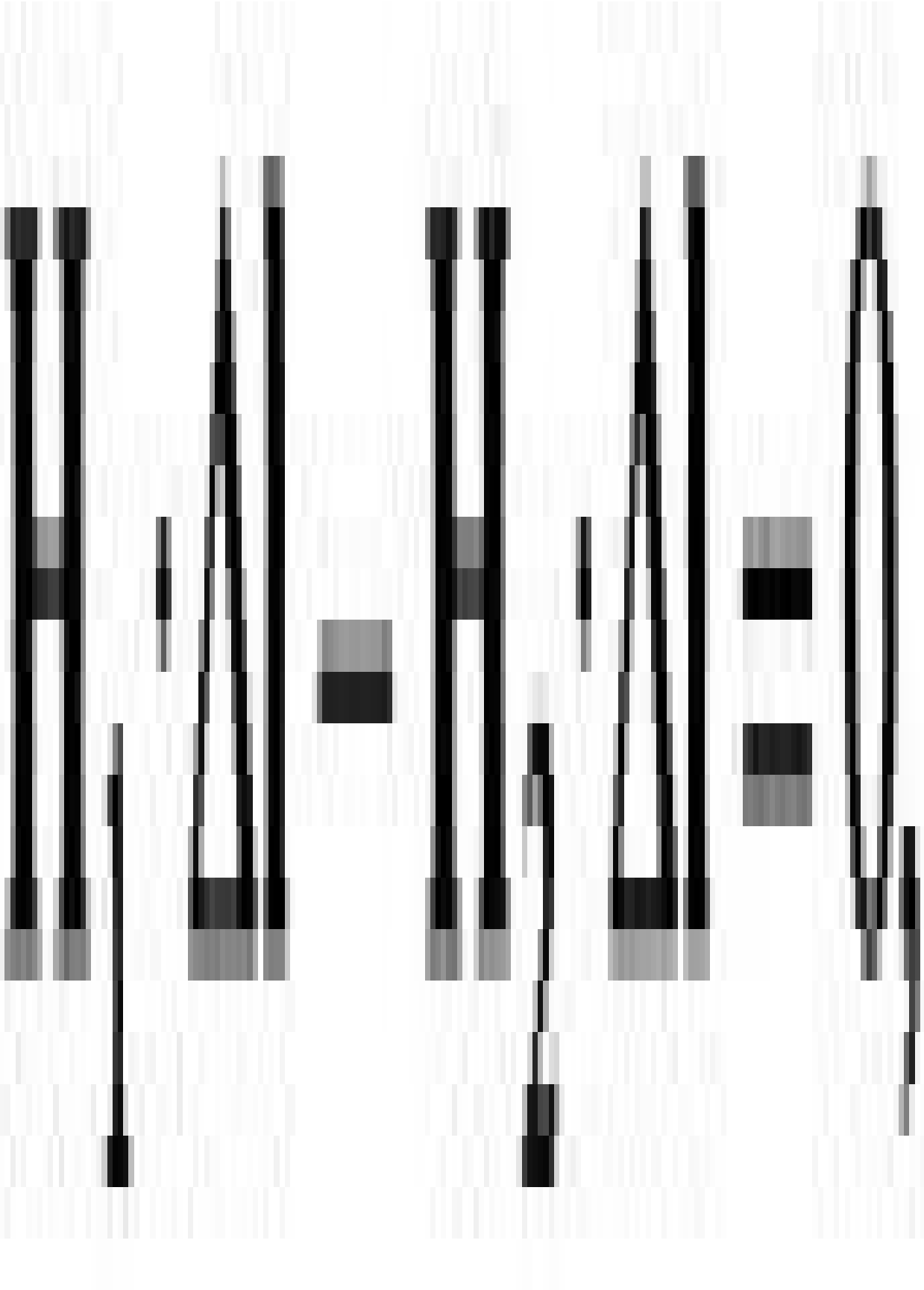
Figure 11.10. Part of a surface separating two media. The component of the field E1 in the surface is along AB. The component of the field E2 in the surface is

along DC.


Surface S
B
Δl
E2
C
A
E1
Δl
D
Field E1 has component E1// in surface along AB
Field E2 has component E2// in surface along DC

(11.50)

which is the current through the surface S in the direction given by the right hand rule together with the sense of traversal of the line integral H·dl around the rectangle. Here jf is the free current density at S. For media of finite electrical conductivity σ the limit (11.50) is always zero since the current density cannot be infinite in real situations involving finite charges. For finite conductivity Equation (11.49) thus reduces to

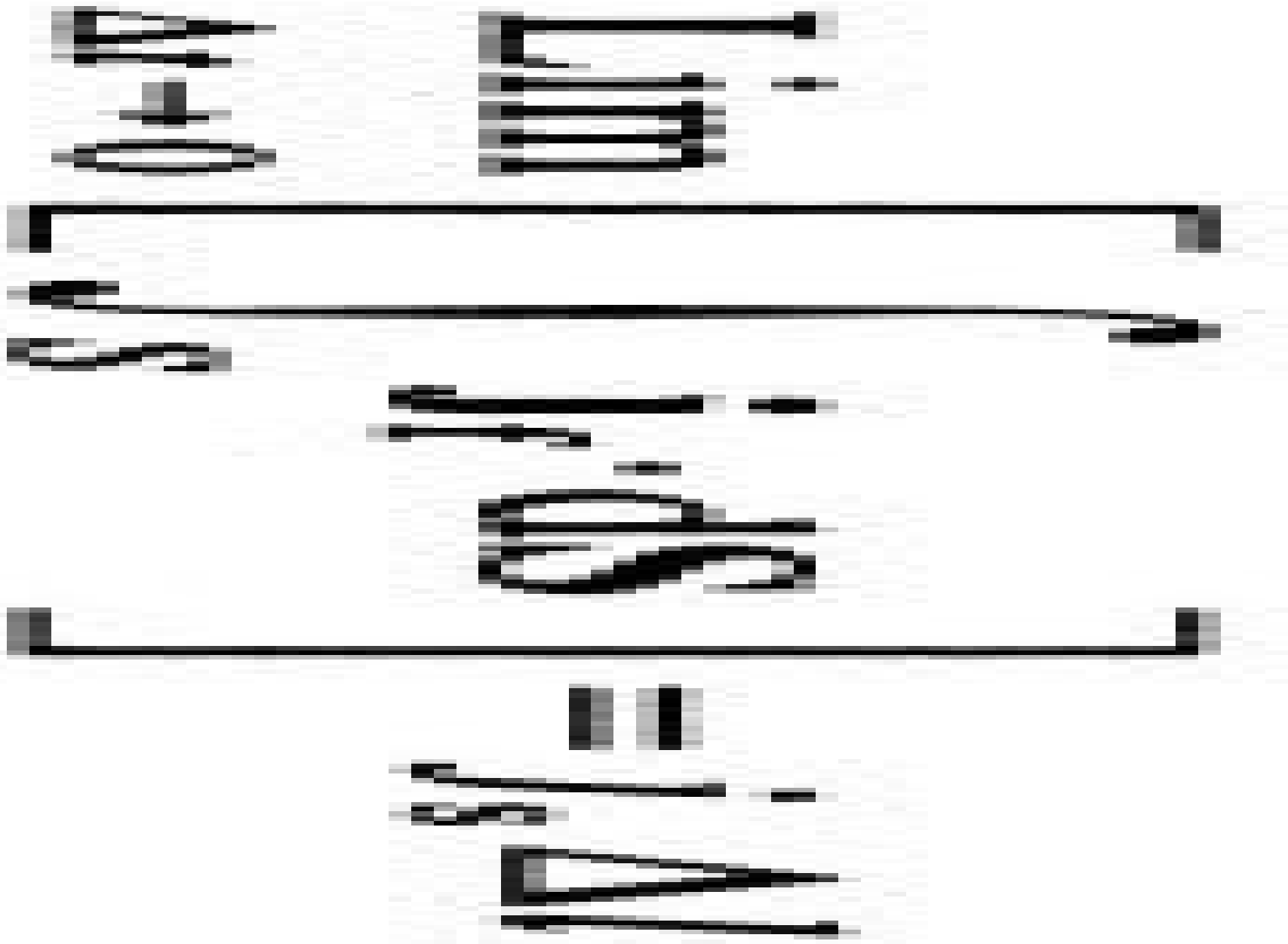


and the boundary condition (as worked out also in section 5.3.2) is that

(11.51)

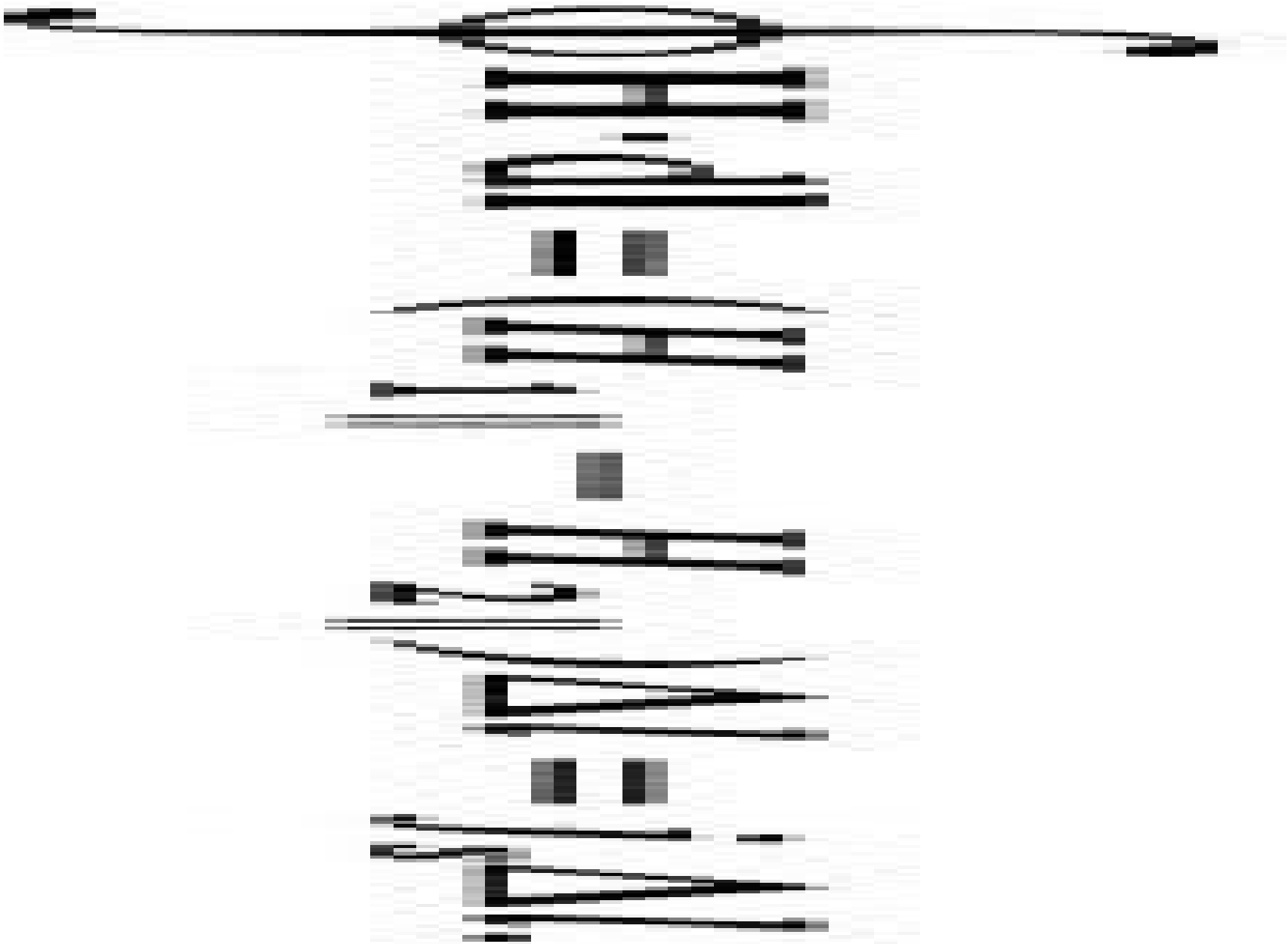
The tangential components of the magnetic intensity are continuous at all points across the interface between two media.

In the treatment of certain problems involving good conductors it is often useful to assume that the conductivities of the conductors are infinite. This assumption greatly simplifies the mathematics, and causes little loss of understanding of the physical processes that occur. With this approximation the limit (11.50) becomes



and Equation (11.49) applied to the loop ABCD can be written

Here js is a surface current per unit length which flows in a vanishingly thin layer, on a macroscopic scale, at the interface between the two media. The current js Δl is again to be taken in the direction given by the right hand screw rule and the sense of traversal of ABCD. There can be no time-varying electromagnetic field inside a perfect conductor since the skin depth is zero. There is thus a discontinuity in the field vector H at the surface. The discontinuity is related to the surface current. Figure 11.11 shows a surface current js flowing in the surface of a perfect conductor. js is perpendicular to the plane of the small rectangular surface ABCD cutting the interface since we have again chosen the sides AB and DC to lie parallel to the projection of the vector H1 on the surface. If we calculate the line integral H·dl around the loop ABCD, traversing the loop from A to B and round to A, we find

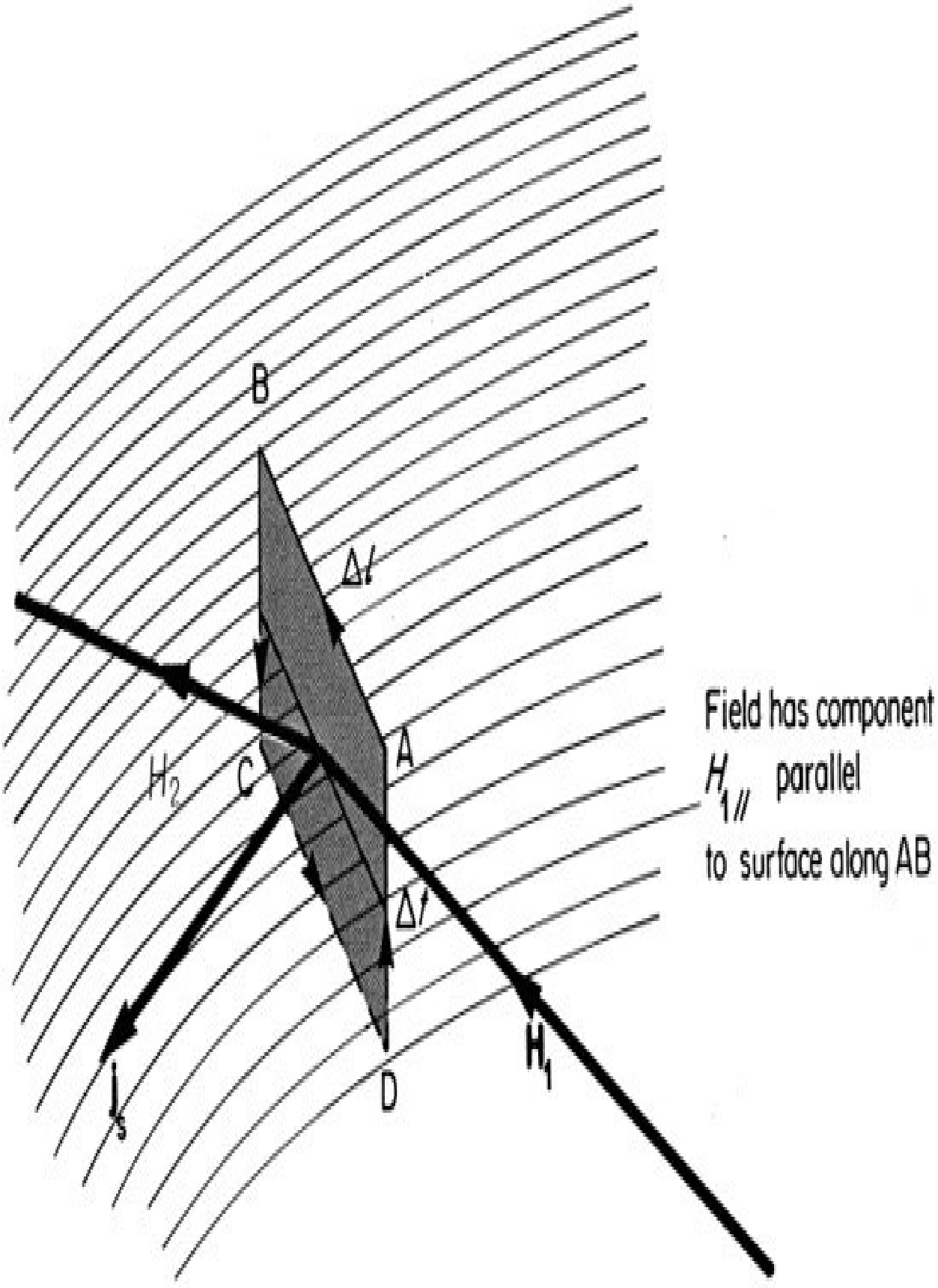


Hence

(11.52)

and at the interface between one medium and a perfect conductor the difference of the components of the magnetic intensity along the surface is equal to the surface current per unit length. This boundary condition is useful for calculating surface currents induced in conductors by electromagnetic waves.

Figure 11.11. Part of a surface separating two media. The lower medium has infinite electrical conductivity, and a surface current is induced in it in order to satisfy the boundary condition (11.52).


B
Δl
A
H2
C
Field has component
H1// parallel
to surface along AB
Δl
H1
D
Is

(iii) The normal component of the magnetic field B

The equation div B = 0 is always true since the lines of the field B are continuous. The argument made in section 5.3.2 is true at all times and the boundary condition is

(11.53)

where $B_\perp$ is the component of the field B normal to the surface.

(iv) The normal component of the field D

The equation div D = ρf, where ρf is the volume density of free charge, leads to the same conclusion obtained in section 2.3.3 that

(11.54)

if there is no surface density of free charge at the interface, and when there is a surface charge density σf,

(11.55)

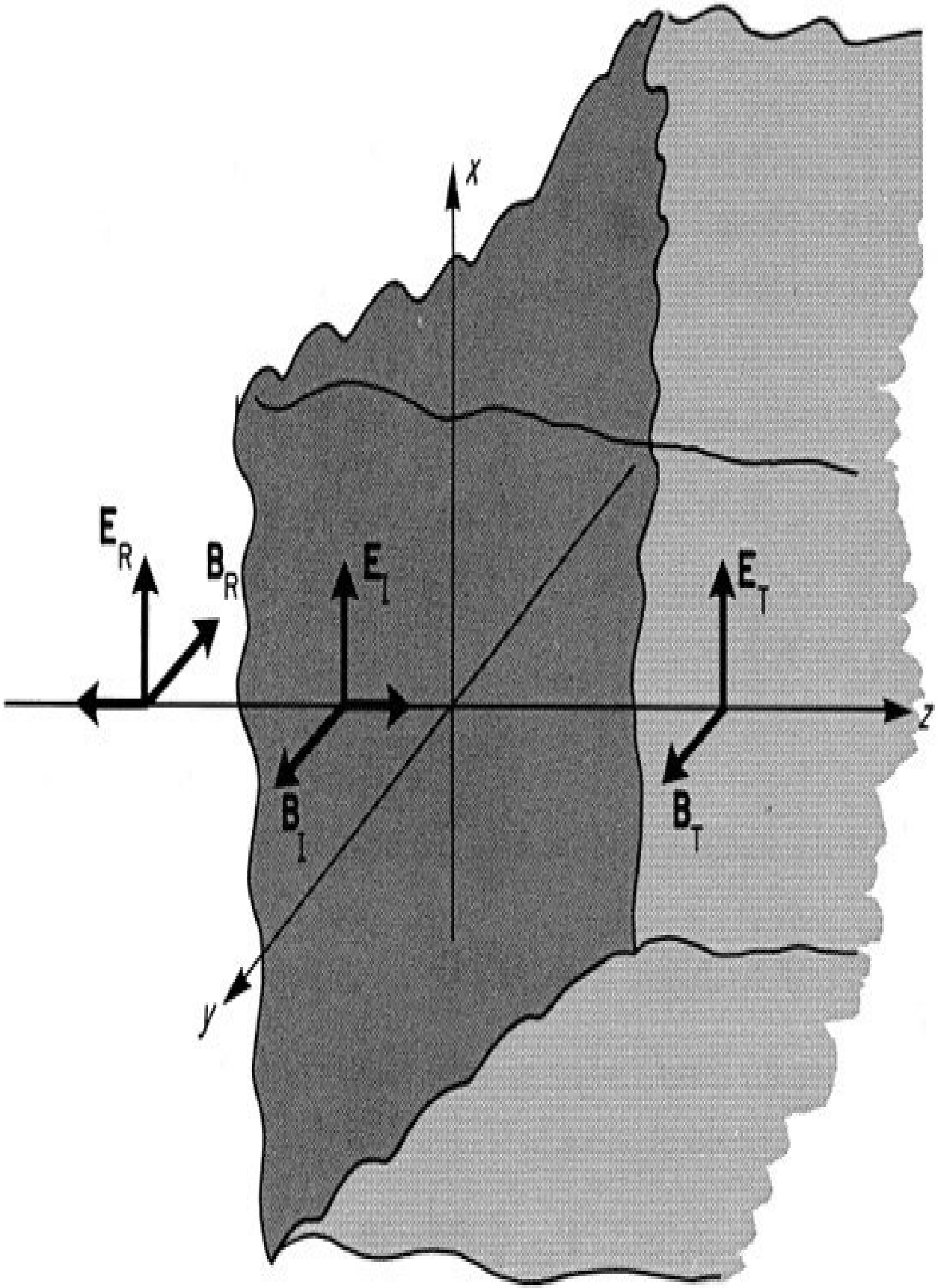
where $D1_\perp$ and $D2_\perp$ are the magnitudes of the perpendicular components of the field D either side of the interface.

11.6.2 Reflection at dielectric boundaries

We will now use the boundary conditions given in the previous section to determine what fraction of the energy in a plane wave incident on a dielectric boundary is reflected, and what fraction is transmitted. We will consider the simplest problem where a plane wave, of angular frequency ω, is incident normally on the boundary. All directions in the plane of the boundary are equivalent in this situation and so we need only consider a single plane polarized wave in order to obtain the answer for a wave of any state of polarization.

Figure 11.12 shows schematically the wave incident on the plane interface. Let the direction of propagation of the wave be the z-direction, and the electric vector lie along the x-axis, so that the magnetic field vectors lie along the y-axis. Let us take the medium on the left to be air or free space, and that on the right to have relative permittivity ε and relative permeability μ. On the left there is an incident wave, whose fields will be denoted by EI and BI, and a reflected wave with fields ER and BR. On the right there is a transmitted wave with fields ET and BT. We can write down the plane wave expressions for these fields:

Figure 11.12. A plane polarized wave incident normally onto the plane boundary between a dielectric on the right and free space on the left. The arrows indicate the relative directions of the fields in the incident, reflected and transmitted waves.


x
z
y
E_R
B_R
E_I
B_I
E_T
B_T

$$\mathbf{E}_I = \mathbf{e}_x E_{0I} \exp\left[j(\omega t - k_1 z)\right]$$

$$\mathbf{B}_I = \mathbf{e}_y B_{0I} \exp\left[j(\omega t - k_1 z)\right]$$

$$\mathbf{E}_R = \mathbf{e}_x E_{0R} \exp\left[j(\omega t + k_1 z)\right]$$

$$\mathbf{B}_R = -\mathbf{e}_y B_{0R} \exp\left[j(\omega t + k_1 z)\right]$$

$$\mathbf{E}_T = \mathbf{e}_x E_{0T} \exp\left[j(\omega t - k_2 z)\right]$$

$$\mathbf{B}_T = \mathbf{e}_y B_{0T} \exp\left[j(\omega t - k_2 z)\right]$$

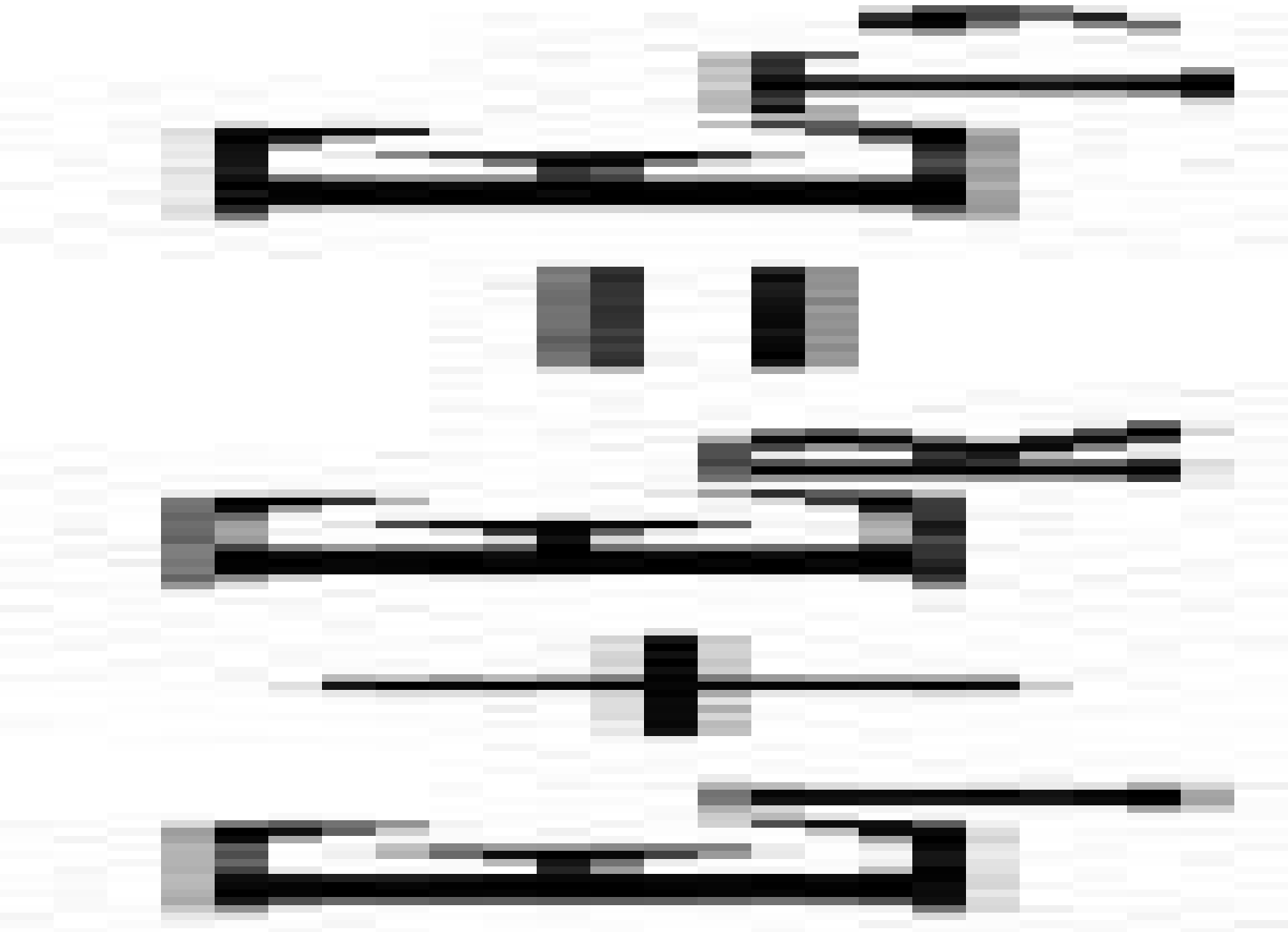
where

$$k_1 = \omega \sqrt{\varepsilon_0 \mu_0}$$

$$k_2 = \omega \sqrt{\varepsilon \varepsilon_0 \mu \mu_0} .$$

In these expressions E0I, B0I, … etc., are the amplitudes of the fields, and we have used the result of Equation (11.25) for the relative directions of the electric and magnetic fields. It might be thought that we should introduce phase changes between the three sets of fields and that the relative phases would be determined in the solution to the problem. However, the only phase changes which can occur are zero or π, otherwise it would be impossible to satisfy the boundary conditions at all times. This can be seen most easily by noting that on the right of the boundary the field is a pure harmonic wave. In order to match fields across the boundary at all times, the total field on the left must thus be a pure harmonic wave in phase with the field on the right, and we have that the phase change between waves are zero or π as stated. Whether the fields are in phase or π out of phase will emerge naturally from the sign relationships between the amplitudes E0I, E0R, etc. The relationships between the amplitudes, and hence the transmission and reflection coefficients, are determined by making the total fields on either side of the boundary obey the conditions discussed in section 11.6.1.

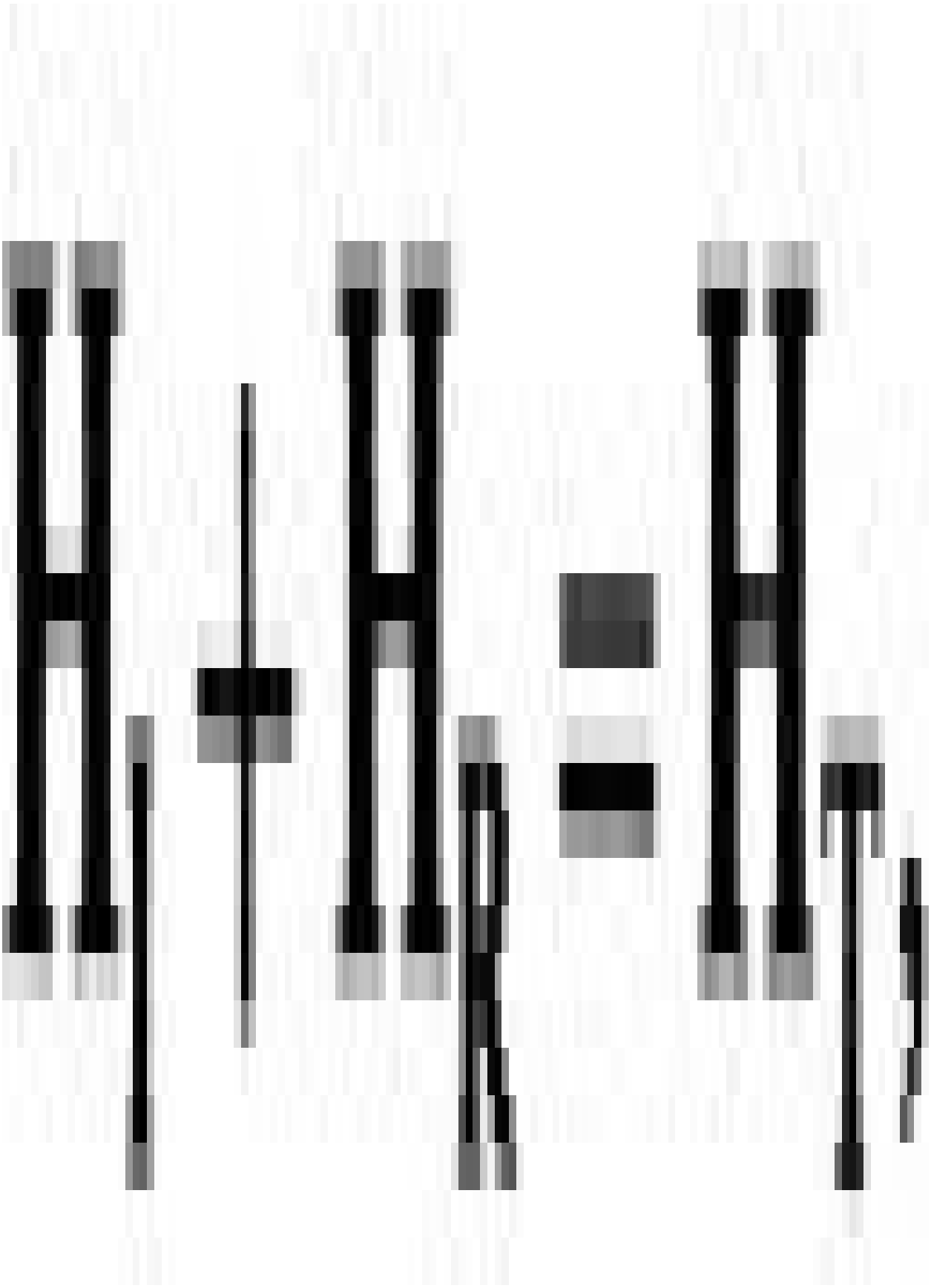
Let us choose the boundary to be at $z = 0$ for convenience. The common factor $\exp(j\omega t)$ can be ignored; if we satisfy the boundary conditions at one time, say $t = 0$, they will be satisfied at all times. The fields in this problem are all tangential to the boundary, and condition (11.48) tells us that



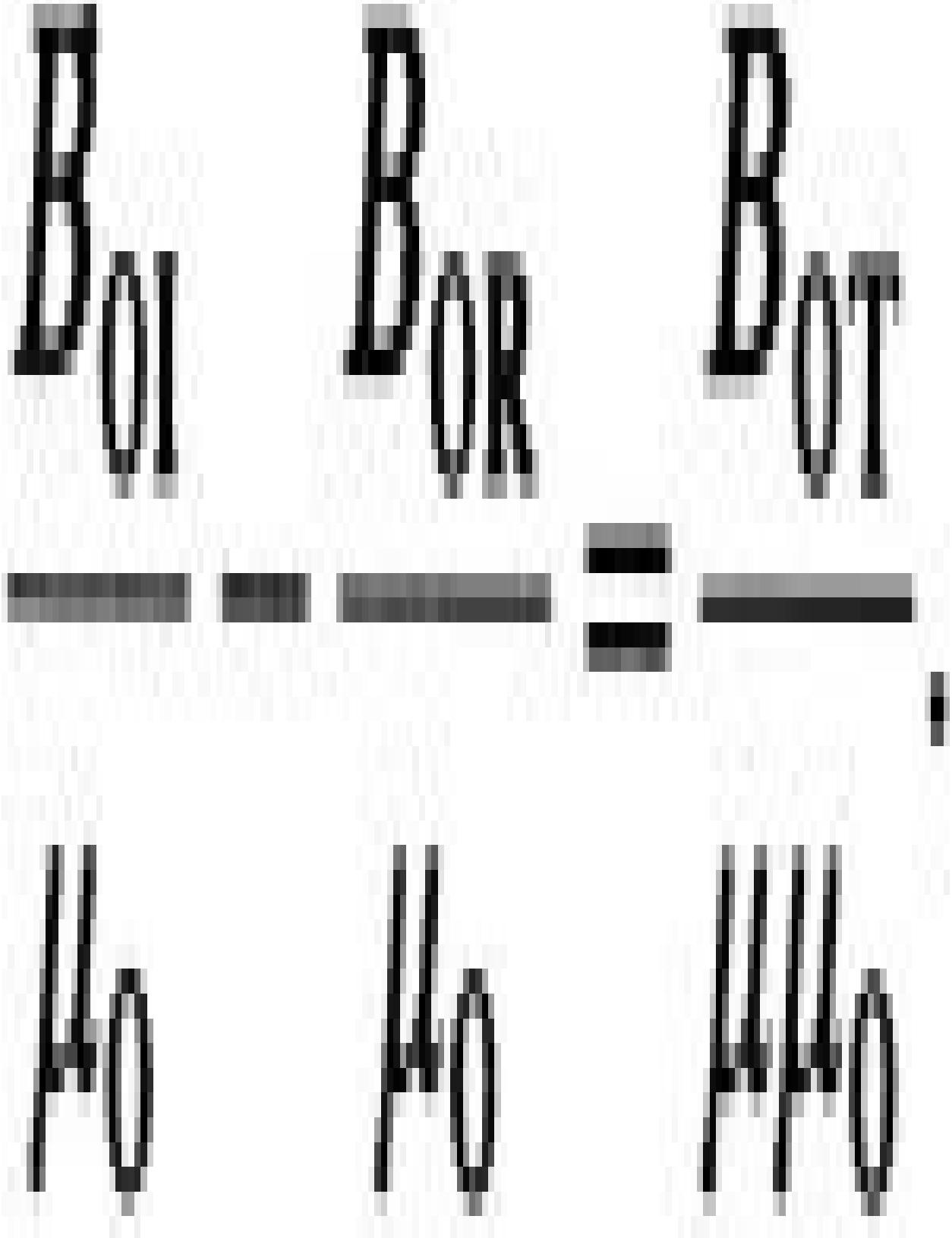
or

(11.56)

There is no surface current at the dielectric boundary and condition (11.51) tells us that



or



Using Equations (11.22) and (11.25) this last equality can be written

(11.57)

Replacing the relative permeability by unity (a good approximation for almost all dielectrics), and using the relation (11.27) between refractive index n and relative permittivity ε, Equation (11.57) becomes

(11.58)

Combining Equations (11.56) and (11.58) we obtain

(11.59)

and

(11.60)

Equation (11.59) gives the ratio of the amplitudes of the reflected and incident waves. For reflection at a medium whose refractive index is greater than unity the ratio is negative. This corresponds to a phase change of π on reflection, i.e. the electric vector of the incident wave oscillates 180° out of phase with that in the reflected wave. This is the phase difference used in interference calculations

in optics when one part of a split wave front is reflected at a medium of greater refractive index than the medium in which it is incident.

The reflection coefficient R is defined as the reflected energy divided by the incident energy. These quantities are given by the time average over a cycle of the Poynting vectors for the incident and reflected waves.

(11.61)

from Equation (11.59).

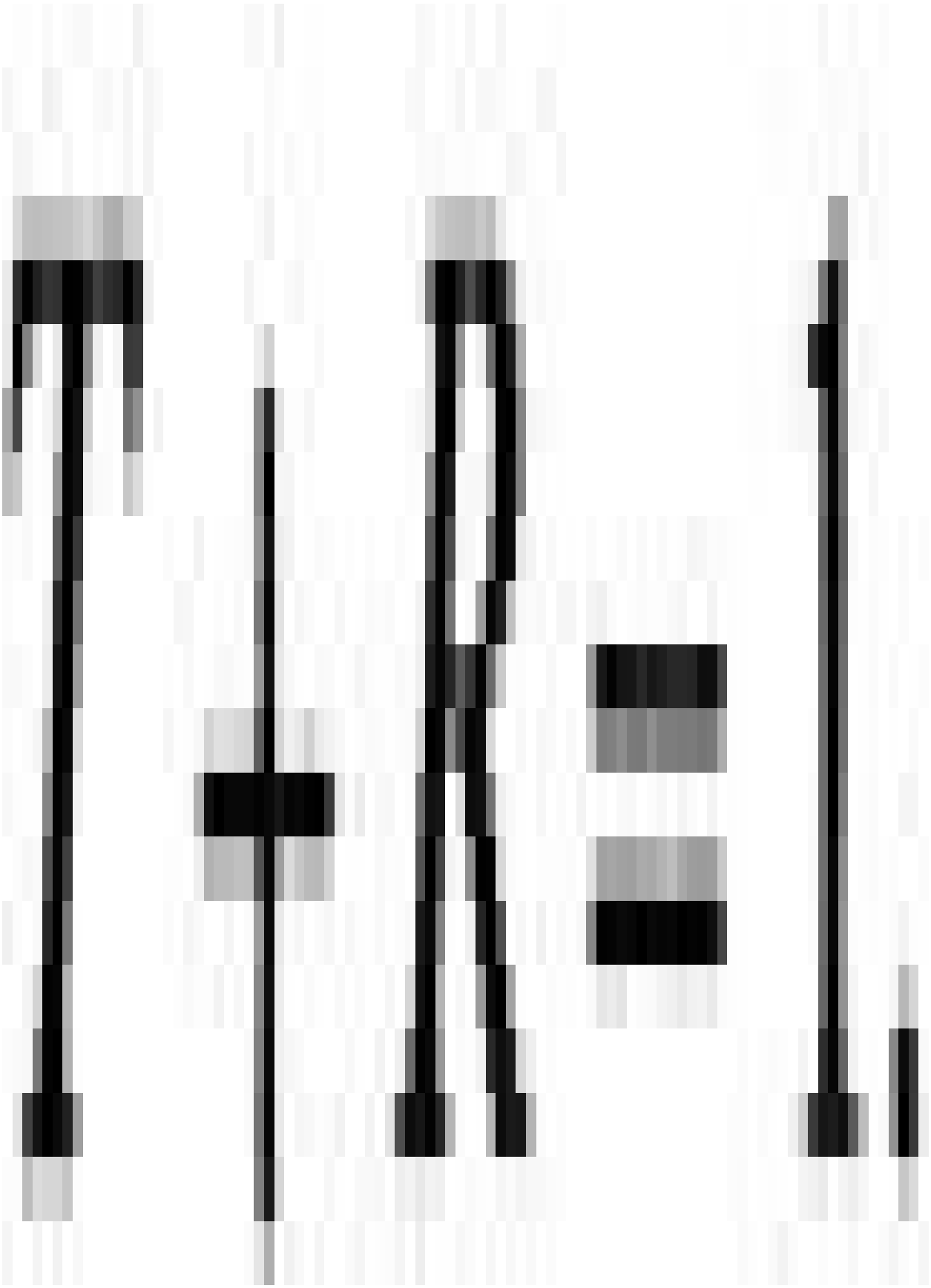
The transmission coefficient T is defined as the energy transmitted divided by the energy incident.

$$T = \frac{E_{0T}^2}{E_{0I}^2} \frac{\sqrt{\varepsilon\varepsilon_0 \mu\mu_0}}{\sqrt{\varepsilon_0 \mu_0}}$$

from Equations (11.22) and (11.25). Again putting $\mu = 1$, we obtain from Equation (11.60)

(11.62)

This could have been deduced also from Equation (11.61), since conservation of energy requires that



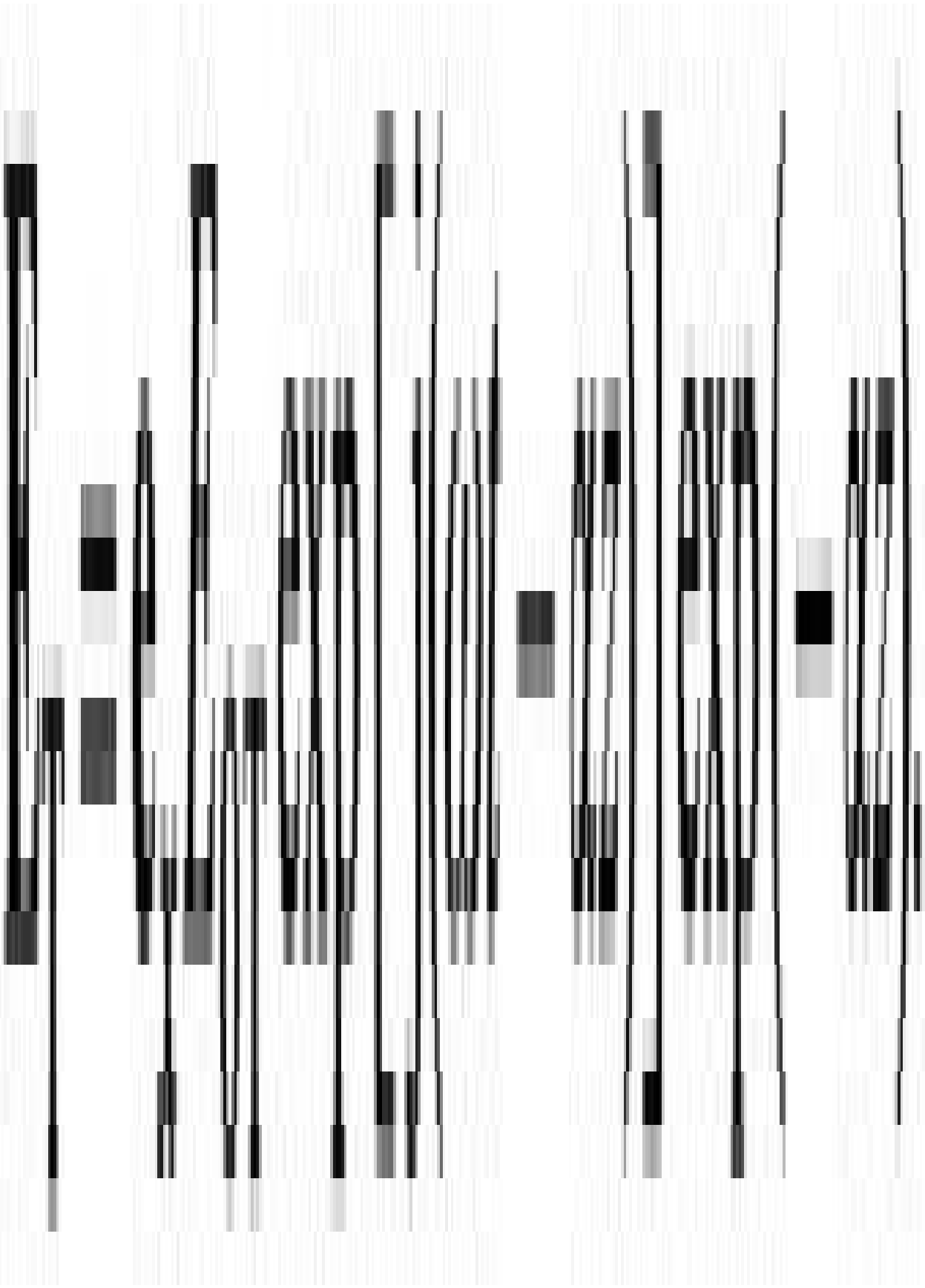
For glass, in the visible region of the electromagnetic spectrum, the refractive index n is about 1.5. The reflection and transmission coefficients, calculated from Equations (11.61) and (11.62), are about 4% and 96% respectively. The refractive index of water for waves of frequency 100 MHz is about 9. The reflection and transmission coefficients are about 65% and 35% respectively. It is seen that the larger the relative permittivity for a given frequency, the larger the fraction of energy reflected. The microscopic explanation of this is in terms of the waves radiated by the molecular charges as they oscillate due to their interaction with the incident wave. A large relative permittivity corresponds to a large oscillating dipole moment in each molecule, and these dipole moments give rise to a large radiated field. The forward wave in this field interferes destructively with the original wave within the dielectric, and gives rise to a small transmission.

11.6.3 Reflection at metallic boundaries

We will now consider the case of the normal reflection of a plane polarized wave at a metallic boundary. The wave is usually almost completely reflected, and for a perfect conductor with infinite conductivity the transmission is zero since there can be no electromagnetic field inside the conductor.

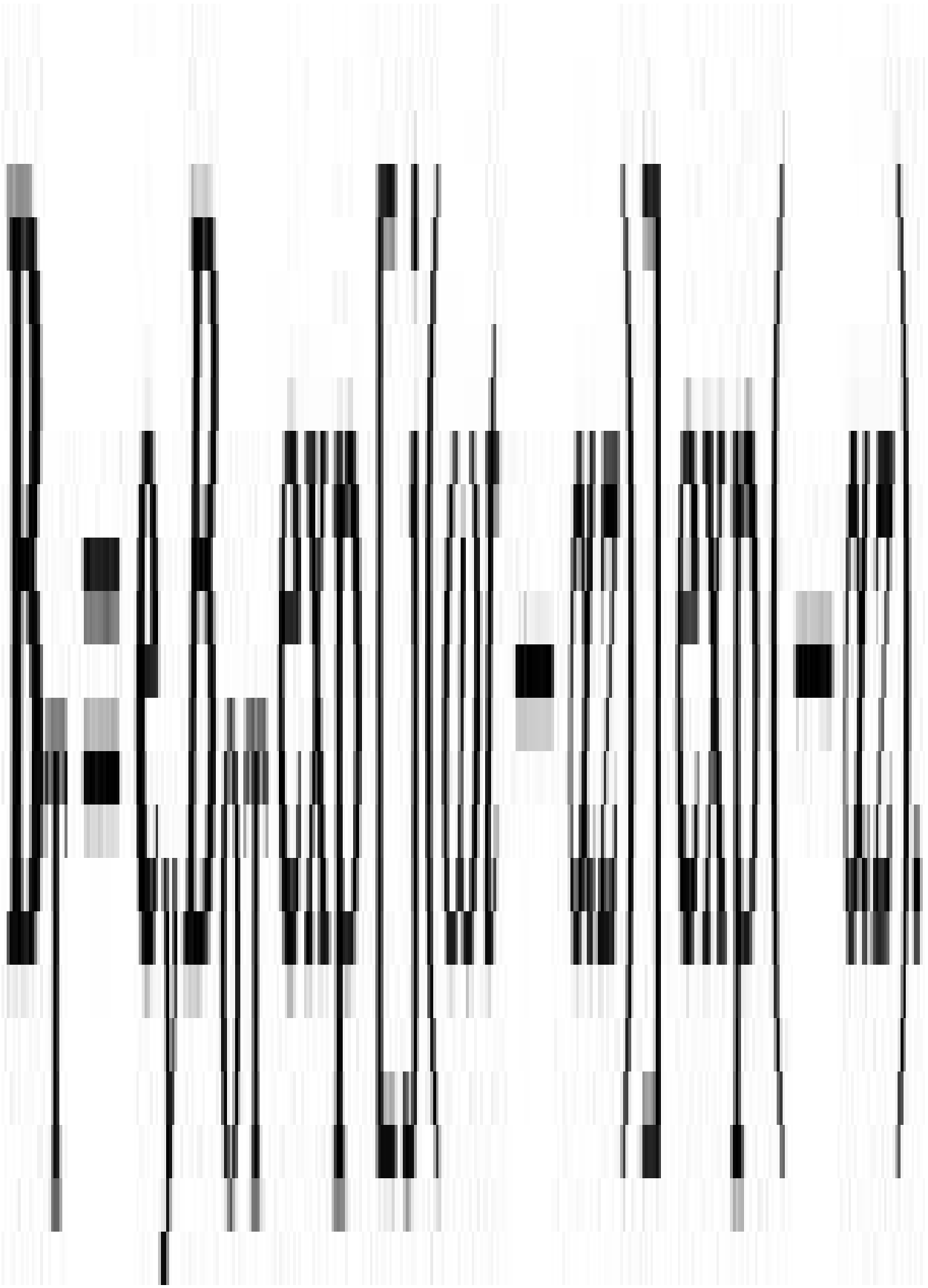
The procedure for determining the reflection and transmission coefficients of a metal of finite conductivity follows that of the previous section. Now, however, since the field inside the conductor is not a pure plane wave but a solution to Equation (11.42), there are phase differences between incident, reflected and transmitted fields. In this section therefore the field amplitudes E0I, E0R, etc., which will be introduced must be considered as complex numbers in order to accommodate these phase changes. The wave in the metal is now an attenuated wave in which the electric field is given by Equations (11.43) and (11.44);

(11.63)



The corresponding magnetic field is

(11.64)



Application of the boundary conditions (11.48) and (11.51) gives the equations

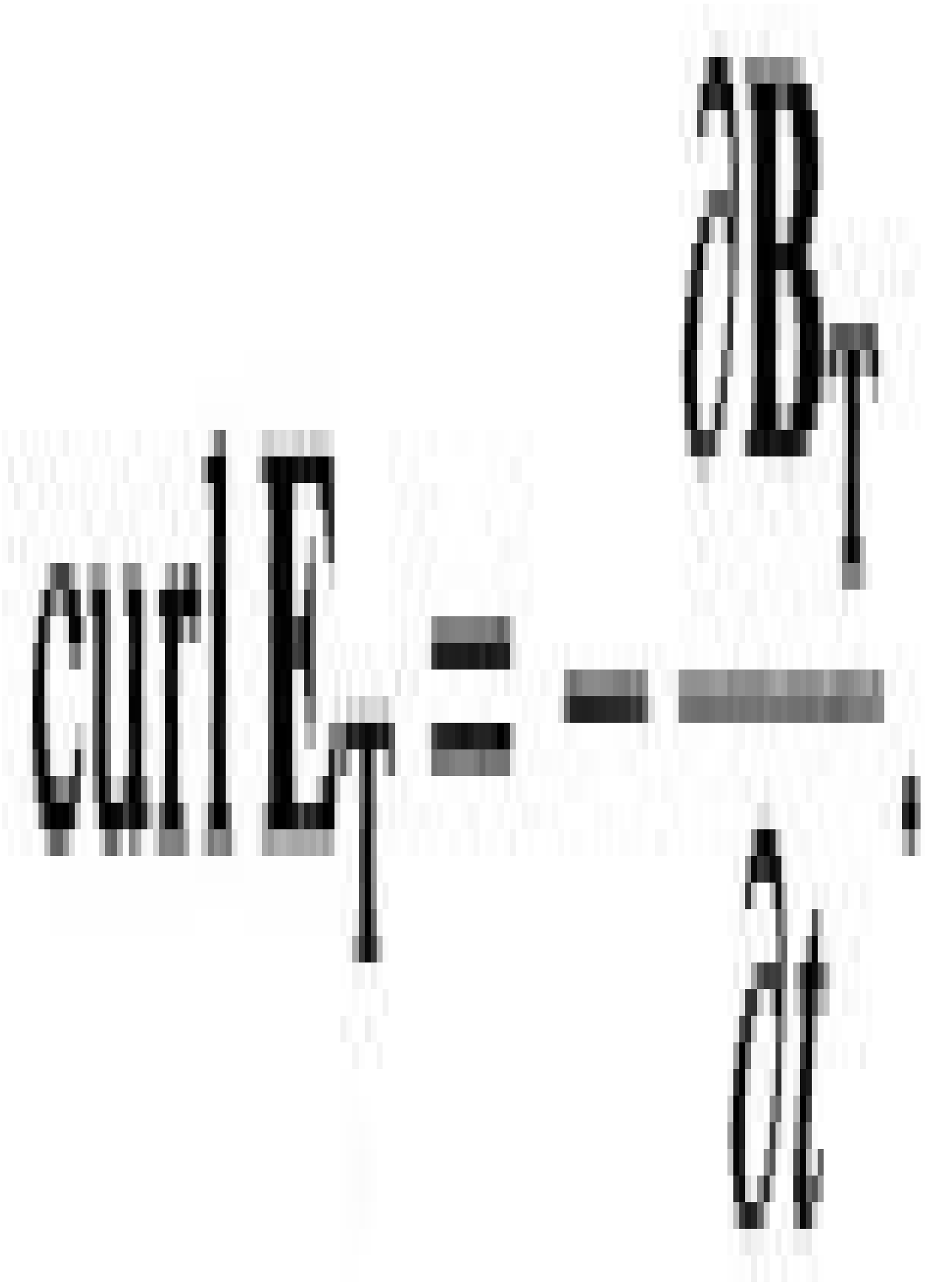
(11.65)

and

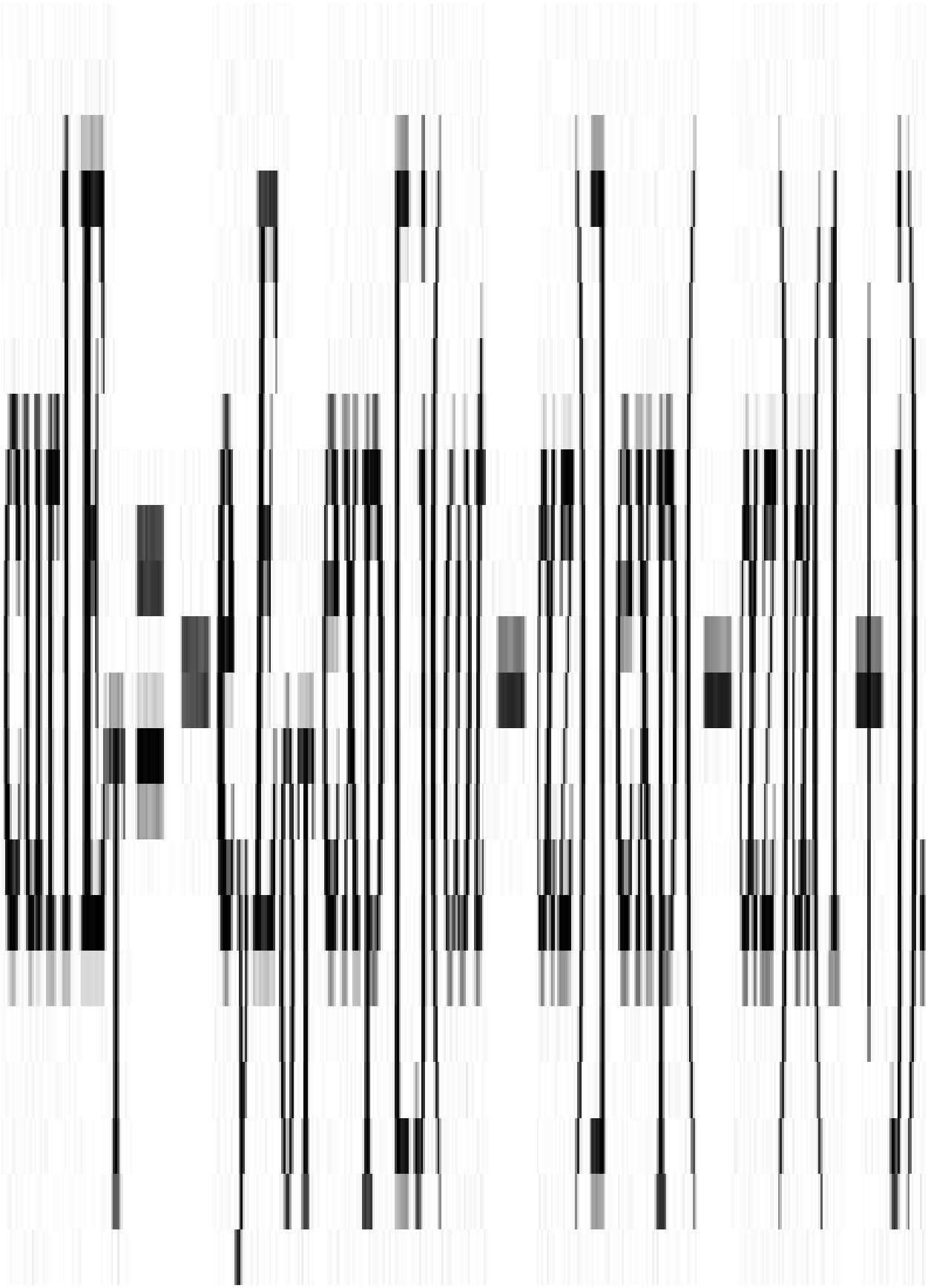
(11.66)

as before. We may use Equation (11.25) again to give us the relation between E0I and B0I, and between E0R and B0R, but we have to return to Maxwell's equations to determine the relation between E0T and B0T.

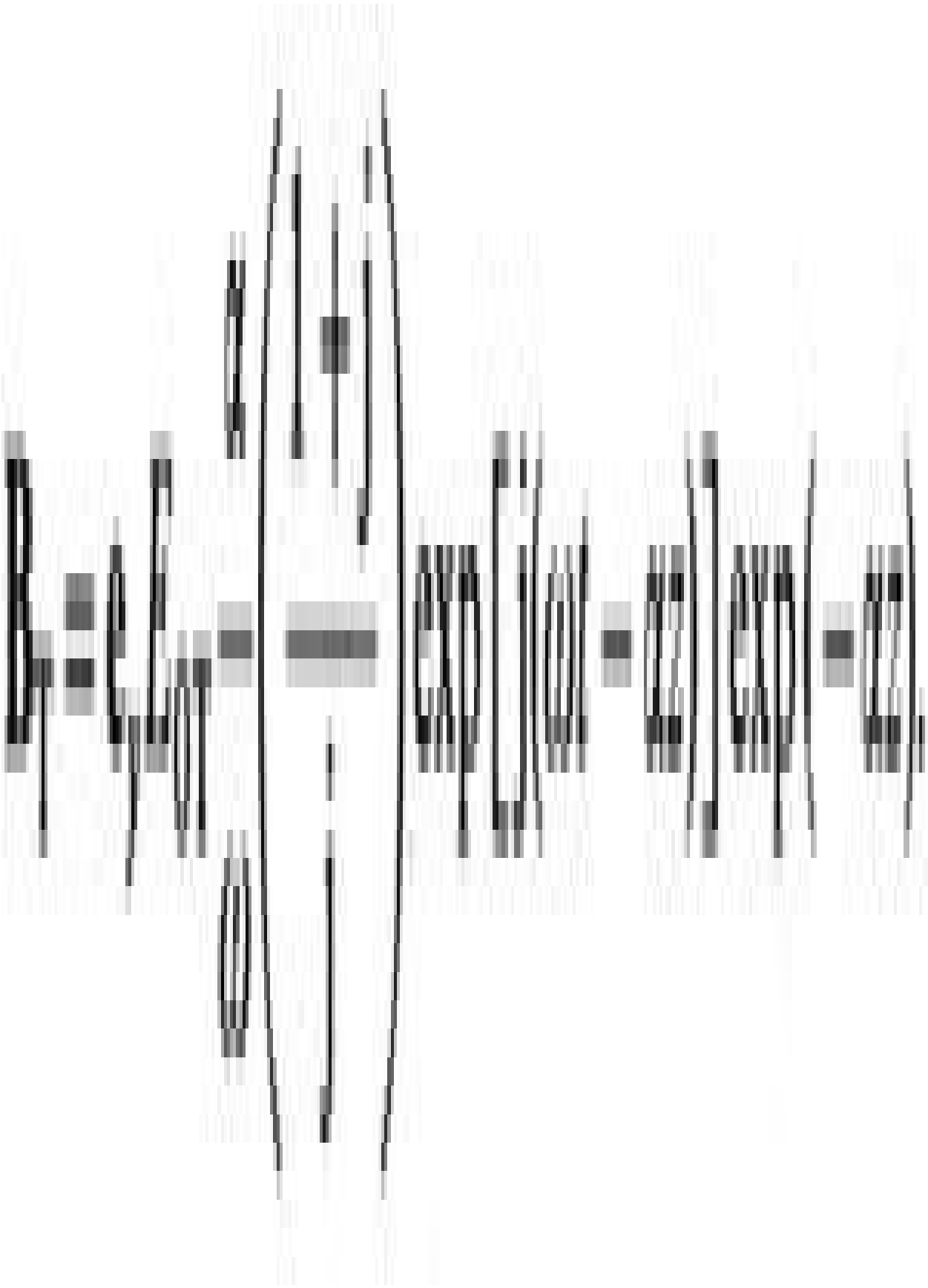
The differential form of Faraday's law holds inside the metal, and everywhere



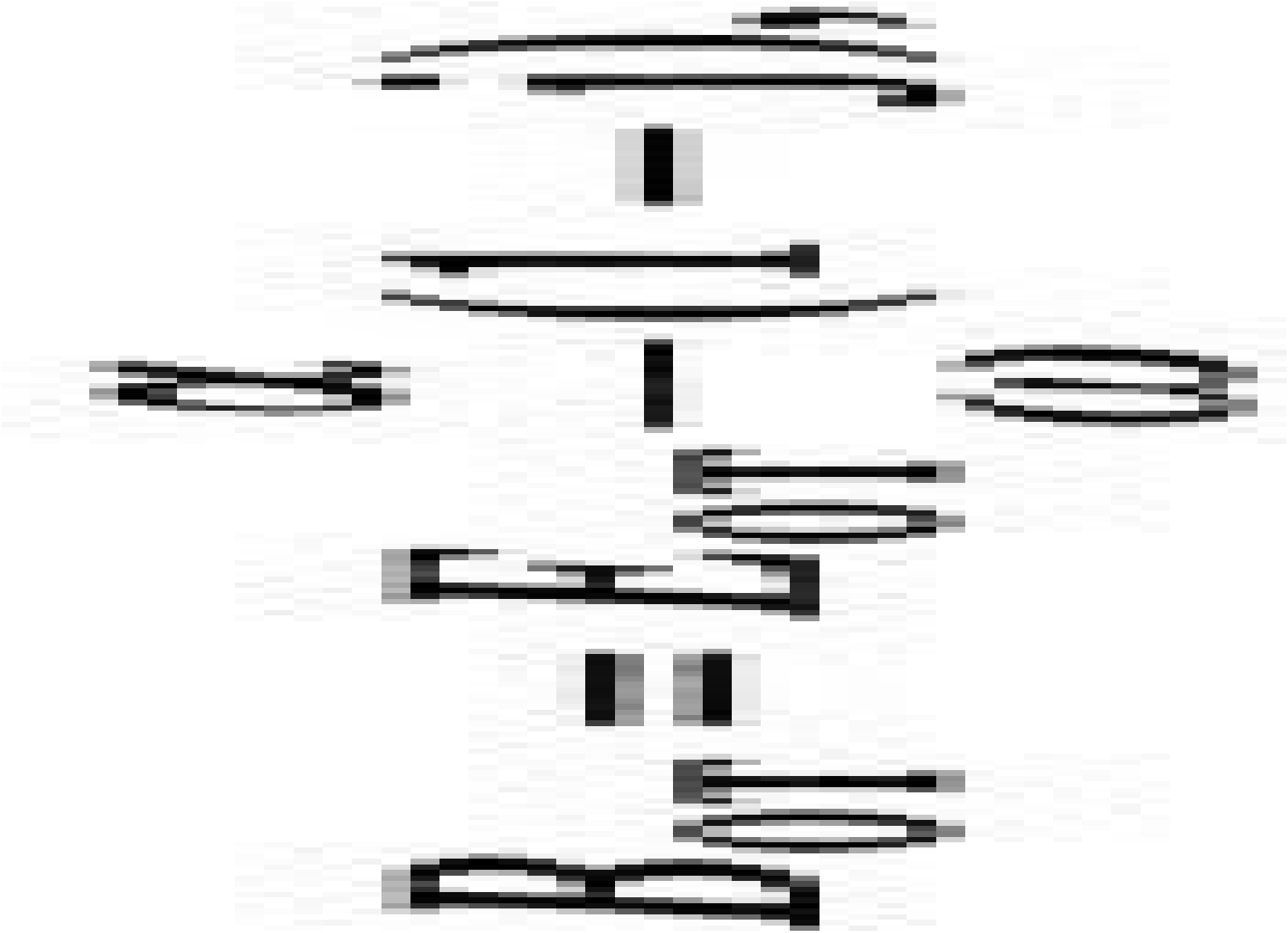
From Equation (11.63),



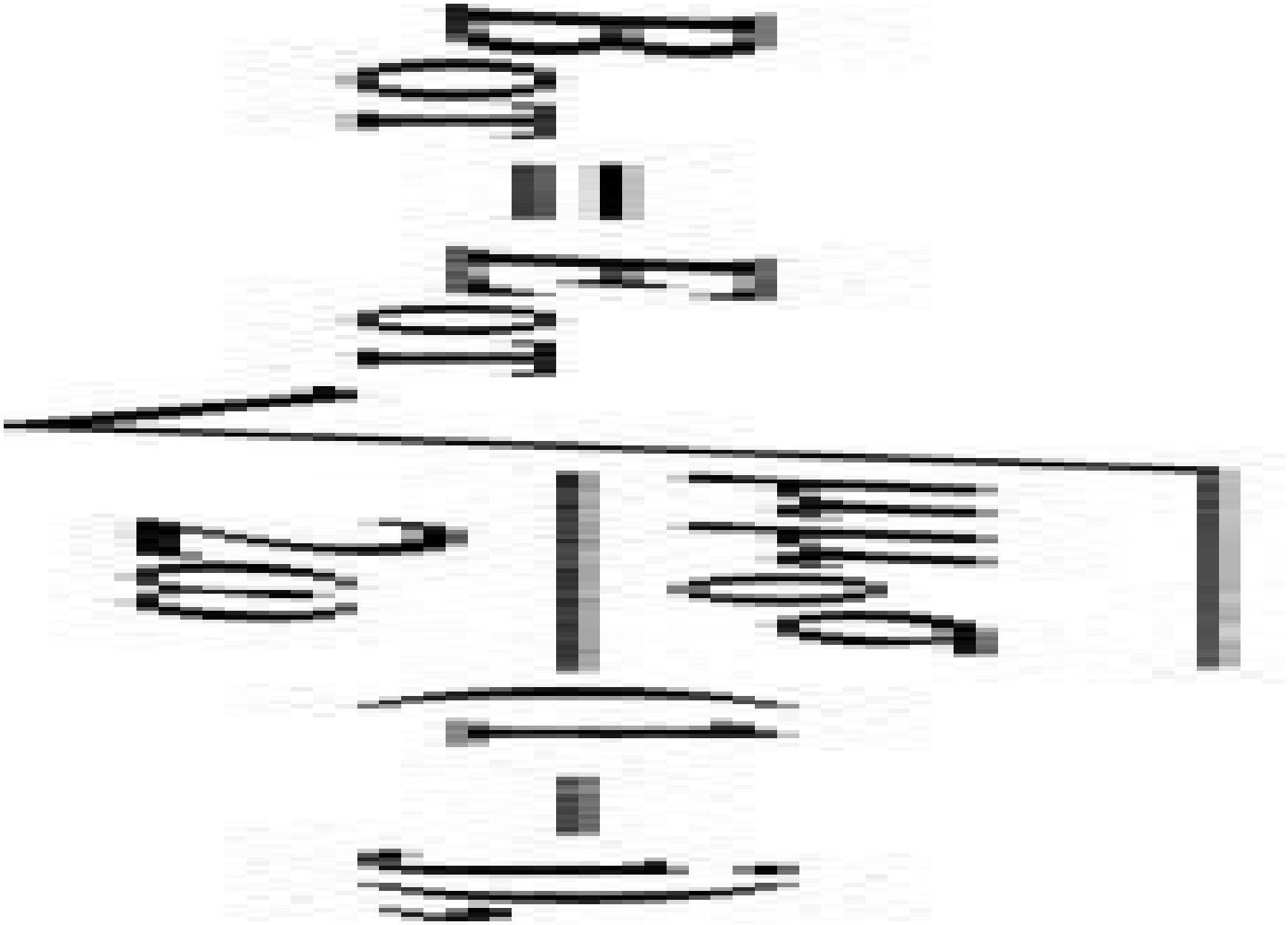
Integration of this equation with respect to time, gives, from Faraday’s law,



We have put the constant of integration equal to zero in the absence of a steady field. Comparison of the above equation with Equation (11.64) shows that



or



from Equation (11.44).

The two Equations (11.65) and (11.66) derived from the boundary conditions can now be expressed in terms of the amplitudes of the electric fields alone. We obtain

$$E_{01} + E_{0R} = E_{0T}$$

$$E_{01} - E_{0R} = E_{0T} \sqrt{\frac{\sigma}{2\omega\mu\varepsilon_0}} (1 - j)$$

Eliminating E0T from these equations, and putting the relative permeability μ equal to unity, gives

(11.67)

The reflection coefficient R is given by

(11.68)

Substitution of the value of E0R/E0l from Equation (11.67) gives

(11.69)

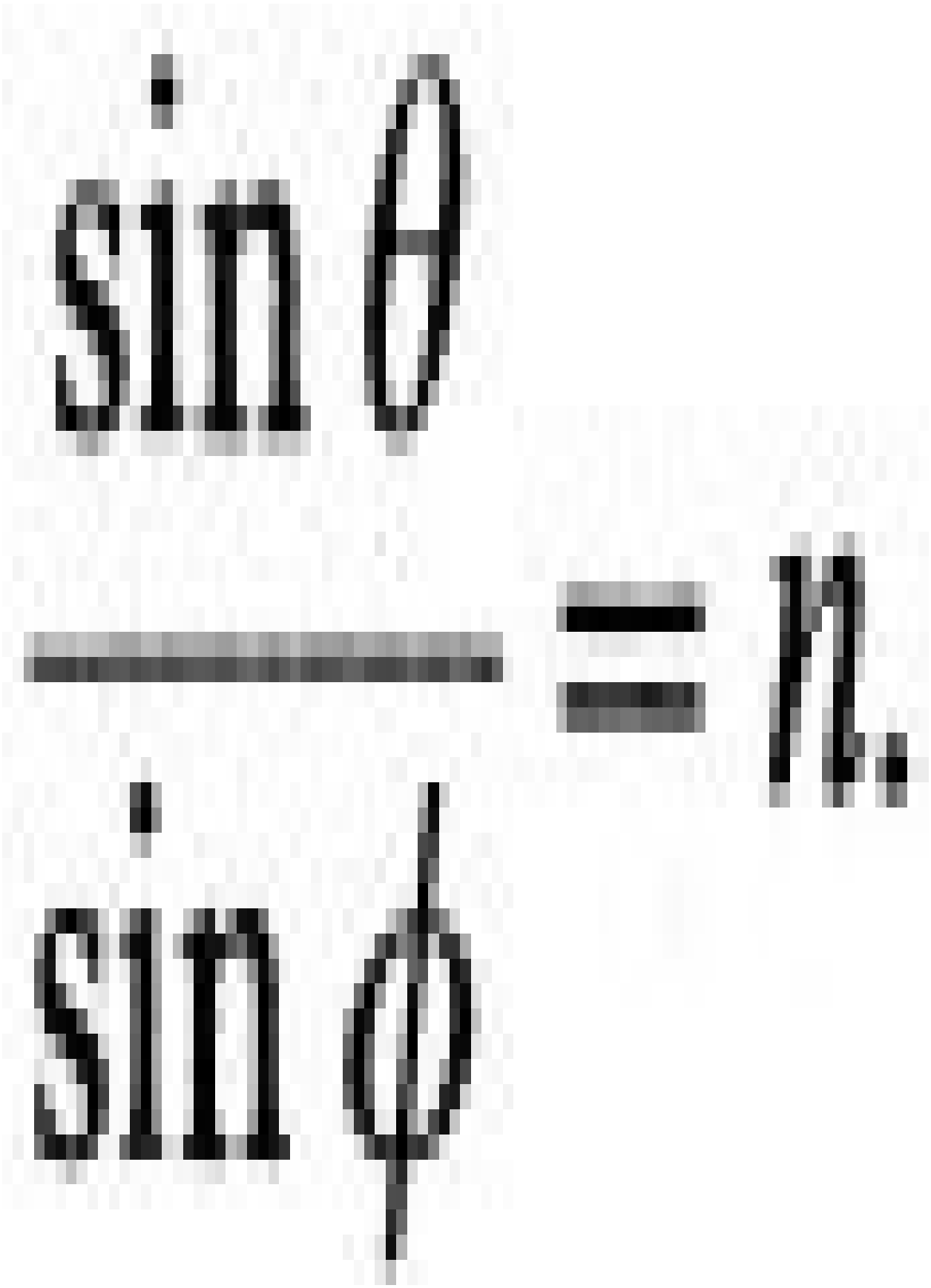
if higher powers of are neglected.

For copper $\sigma \sim 6 \times 10^7$ (ohm m)−1, and at infrared frequencies is equal to about 0.01. The hot filament of an electric heater at 1000 K has an emission spectrum with a peak around a wavelength of $2.5 \times 10-6$ m. A metal reflector behind the filament will reflect about 97% of these infrared rays. The other 3% passes across the metal air boundary but is rapidly absorbed in the metal due to Joule heating. At lower frequencies the reflection is greater and clean metal plates reflect all but a few parts in a million of radio waves incident upon them. At optical frequencies classical theory cannot describe metallic reflection; atomic transitions take place, and these account for the colours of metals.

11.6.4 Polarization by reflection

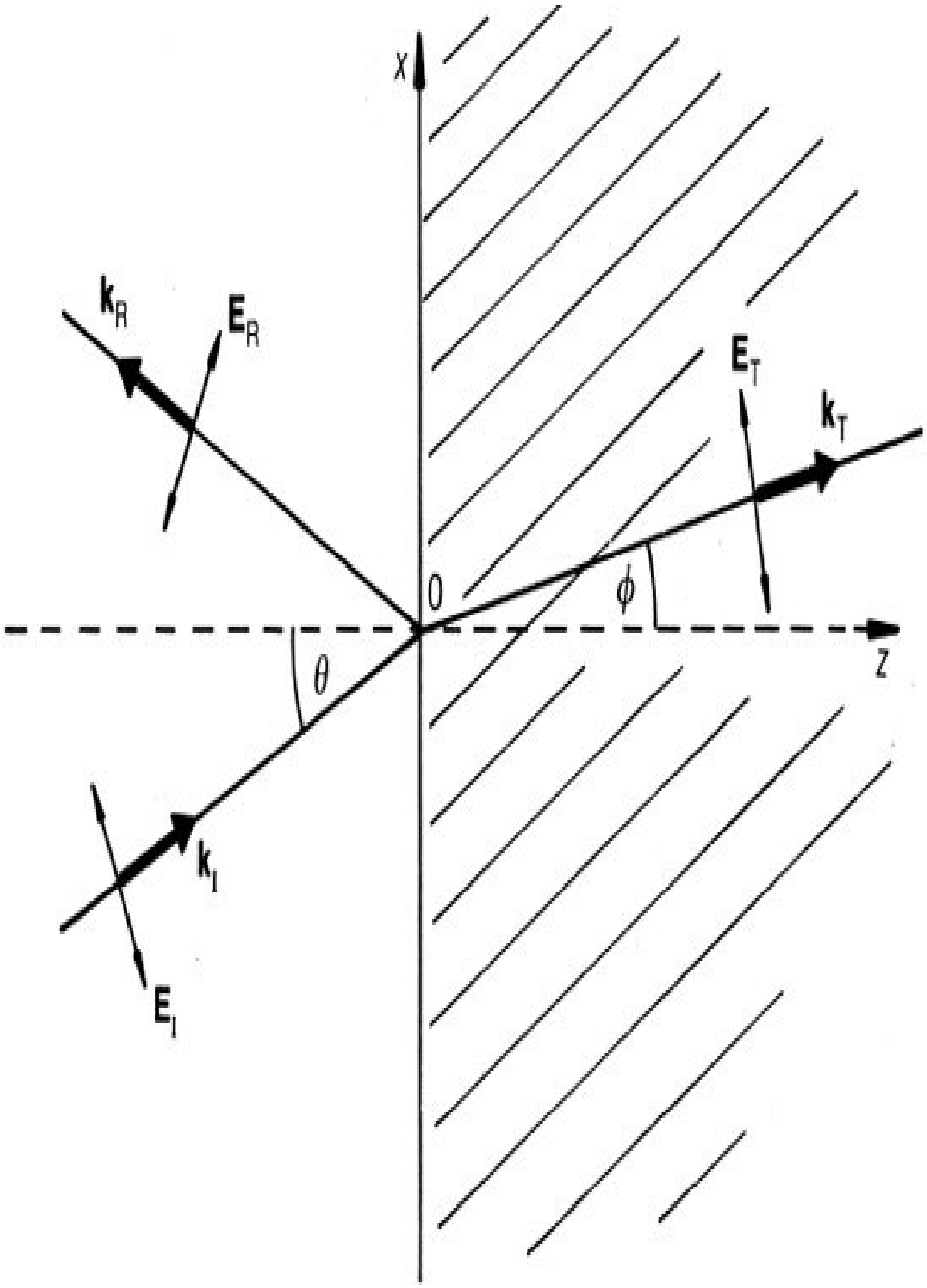
In this section we will consider one way in which a plane polarized wave can be obtained from a wave that is initially randomly polarized, i.e. has its electric field vector changing randomly in the wave front. A randomly polarized plane wave can always be expressed as the sum of two waves which have their electric fields oscillating at right angles to each other. If we can find a process that affects the field in one of these directions differently to the way in which it affects the field in the other we may be able to produce a wave with some degree of plane polarization. If we could remove the field in one direction altogether we would produce a completely polarized plane wave. When radiation strikes a dielectric surface obliquely the reflection coefficient depends on the state of polarization of the incident wave. For plane waves polarized with their electric vectors lying in the reflection plane we will show that for one particular angle of incidence, called the Brewster angle, the reflection coefficient is zero. Hence, if randomly polarized light, such as that from a sodium vapour lamp, is reflected at the Brewster angle, plane polarized light will be generated with the electric field vector perpendicular to the reflection plane.

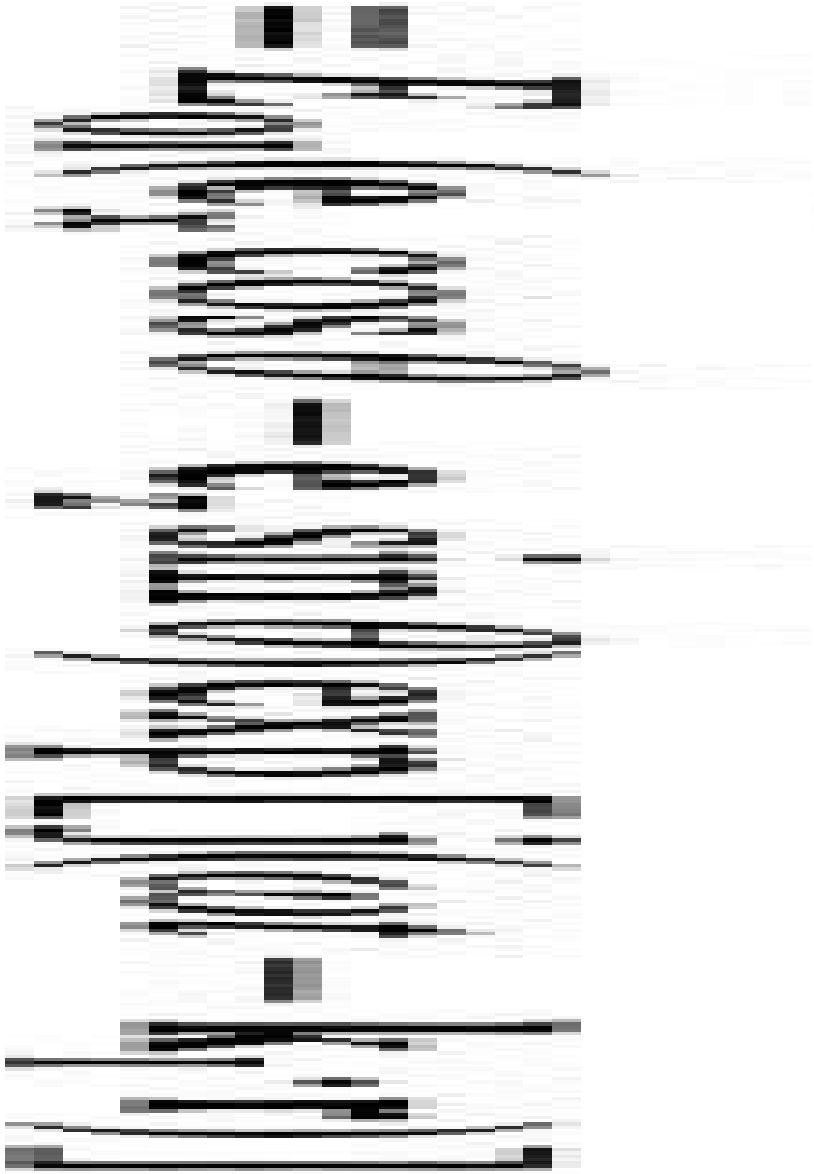
Consider Figure 11.13, in which is shown a view looking down on the boundary between air or free space and a medium which has relative permittivity ε, relative permeability equal to one, and refractive index n. The plane wave is incident at an angle θ to the normal, and the transmitted wave is at an angle ϕ to the normal, where, according to Snell's law,

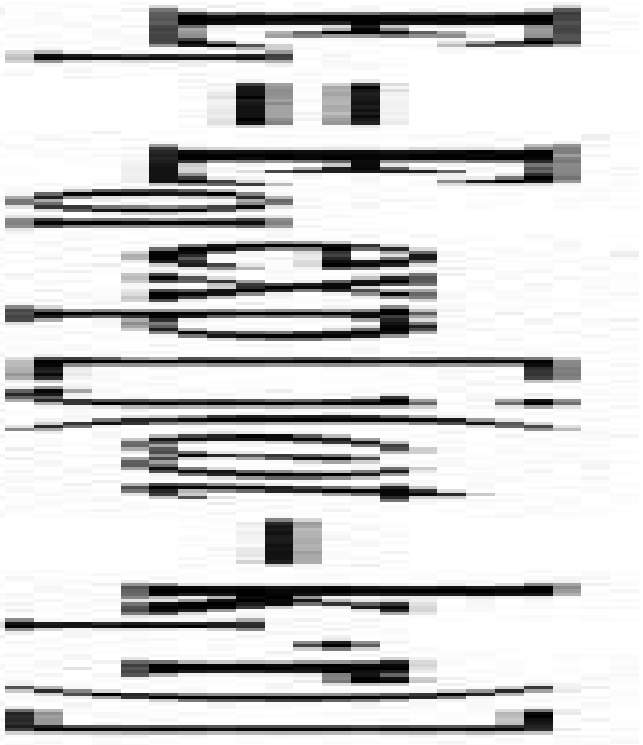


Let us resolve the electric field vector into two components, one in the plane of the paper in Figure 11.13, and the other perpendicular to this. Consider the component E1, in the former plane. With the choice of coordinate system shown on Figure 11.13, the vector E1, lies in the x–z plane and oscillates in directions perpendicular to the propagation vector kI The field EI is given by

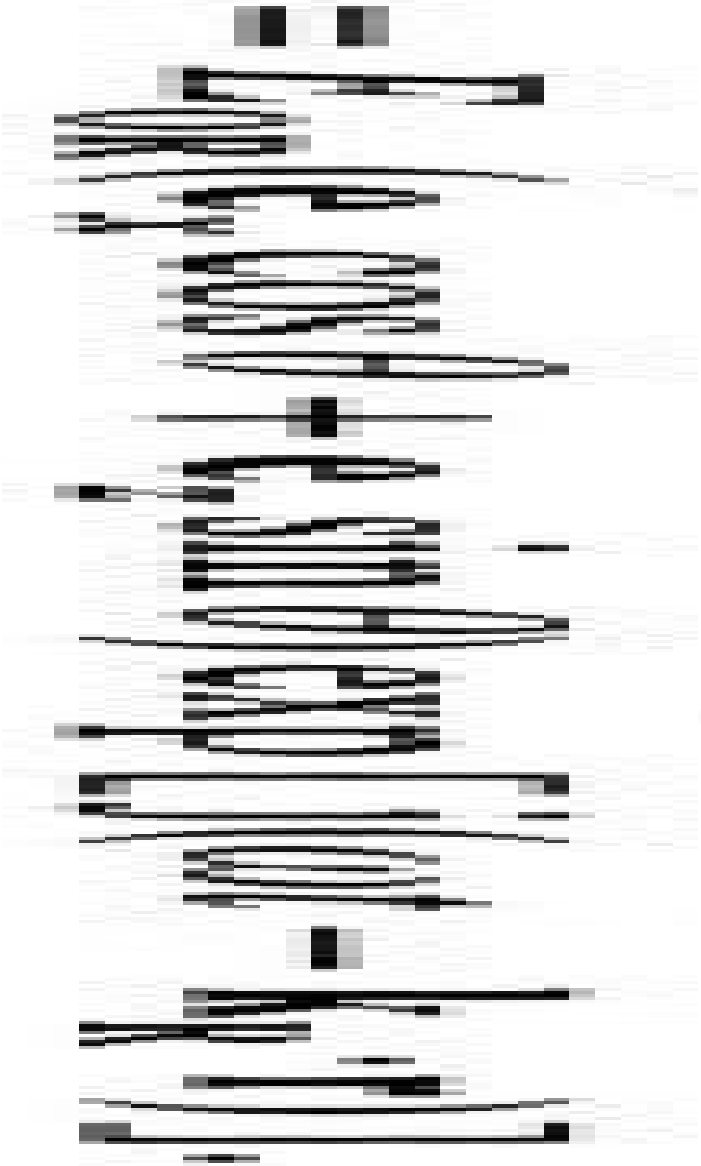
Figure 11.13. The reflection of a plane wave incident obliquely on the plane boundary between a dielectric and free space. The wave is linearly polarized, with the electric fields in the plane of incidence.

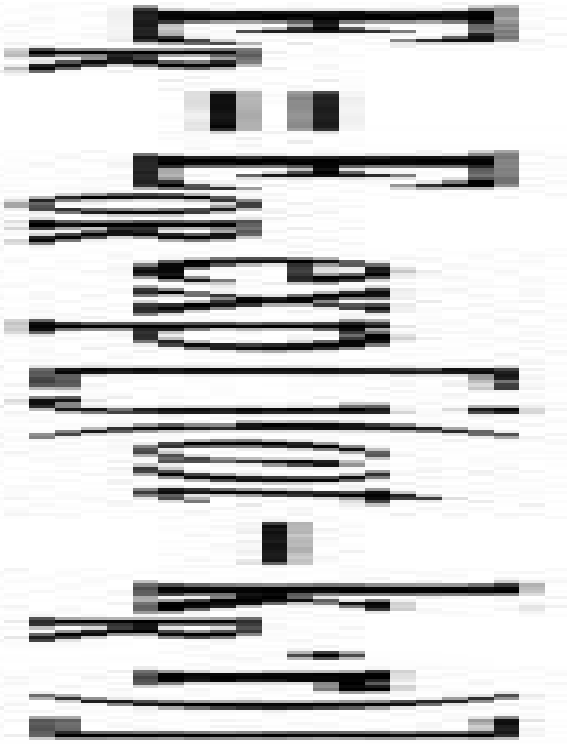

x
k_R
E_R
E_T
k_T
0
ϕ
z
θ
k_I
E_I



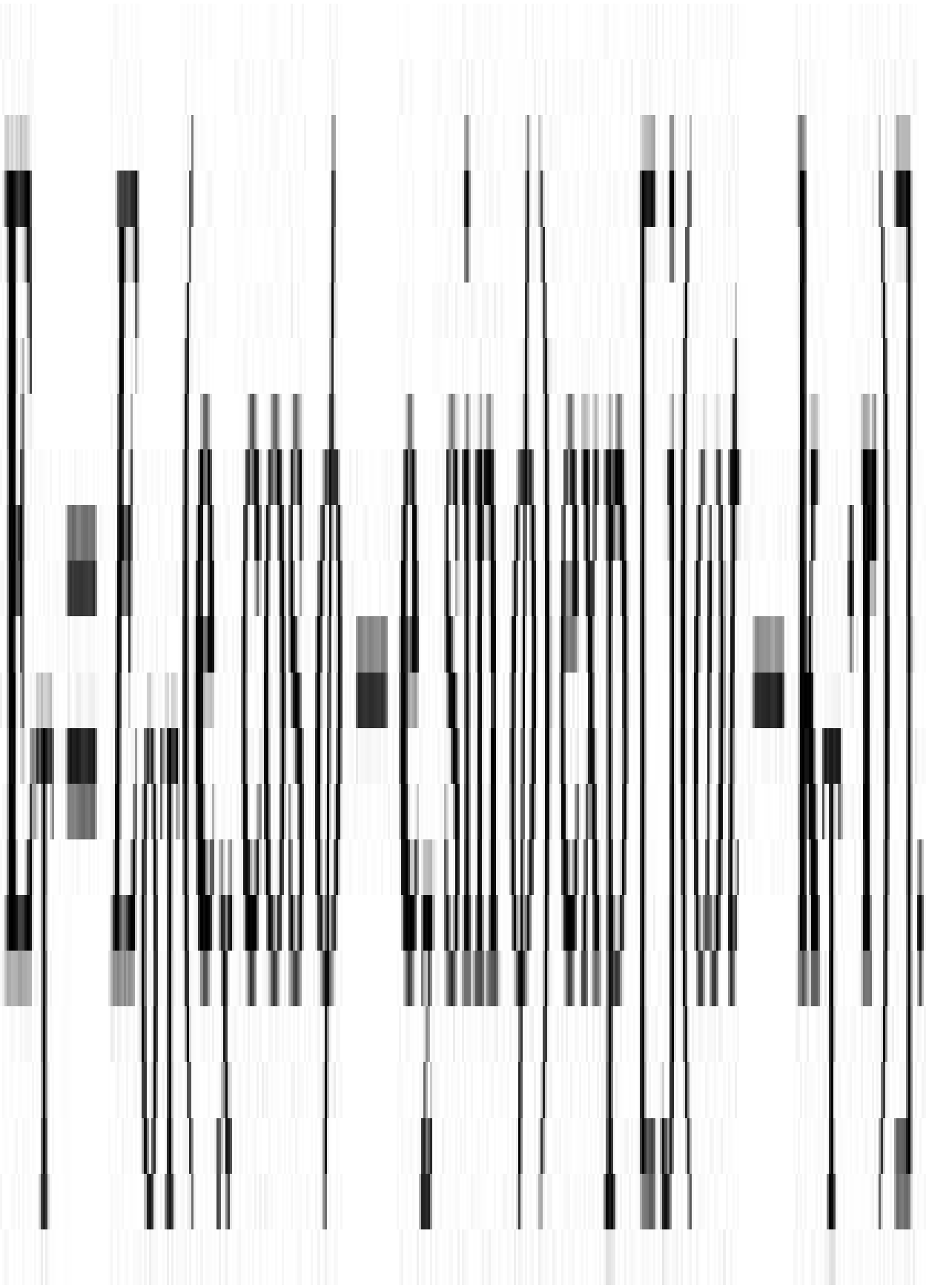


where E0I is the field amplitude. The reflected wave ER has its electric field vector oscillating in directions perpendicular to the propagation vector kR, which is equal in size to kI, but in a different direction, and is given by

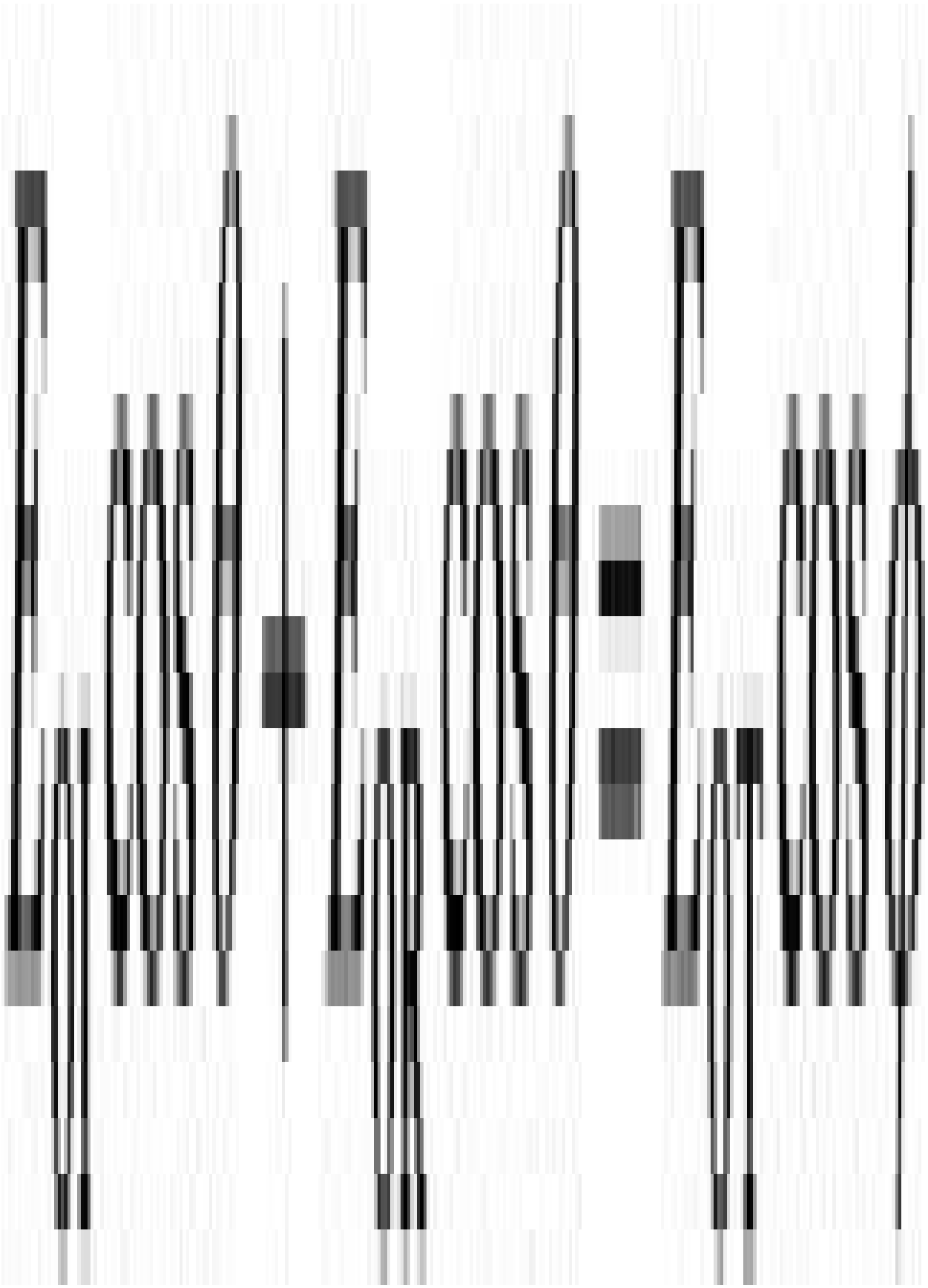




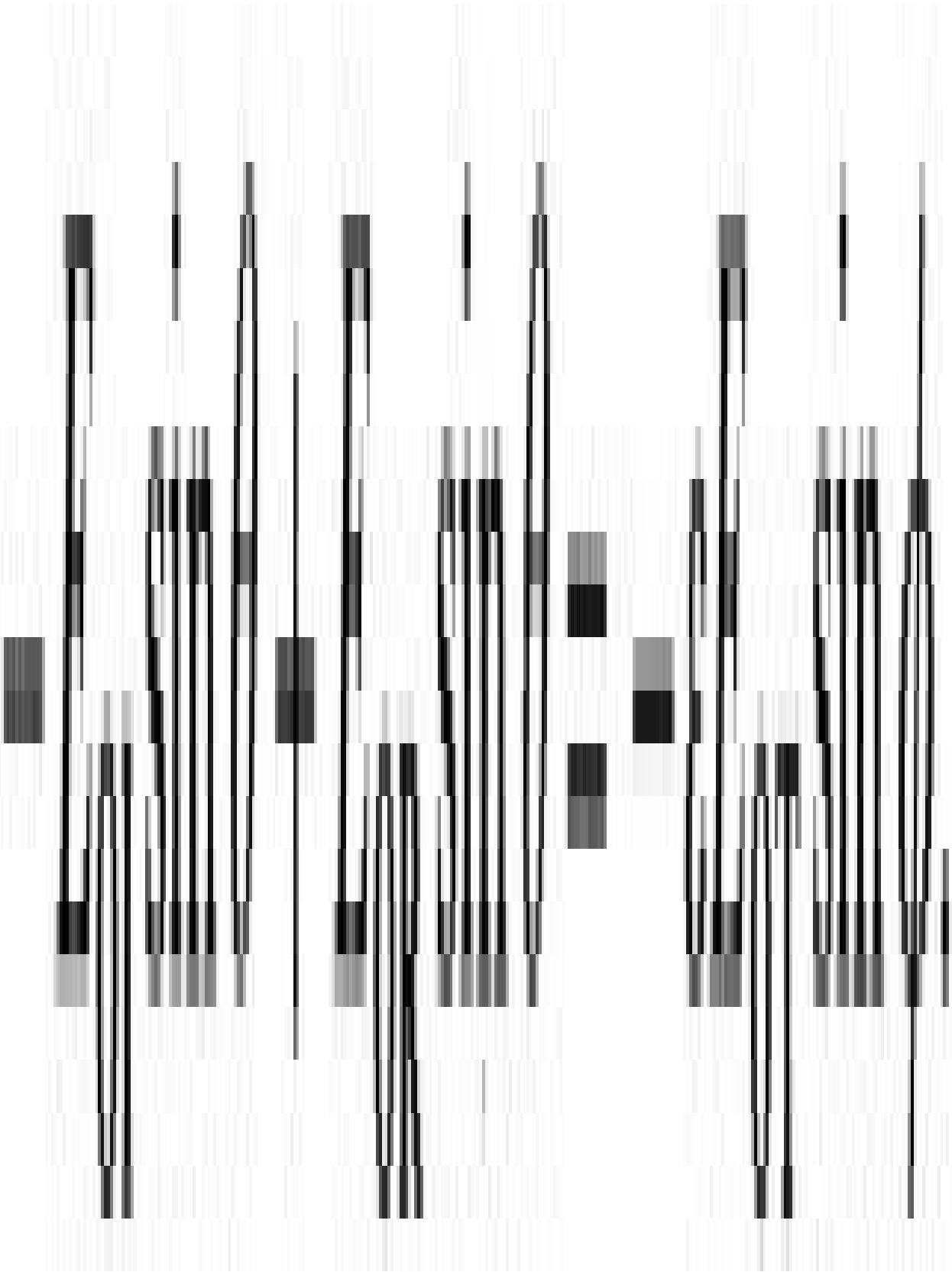
Similarly the transmitted field ET is given by



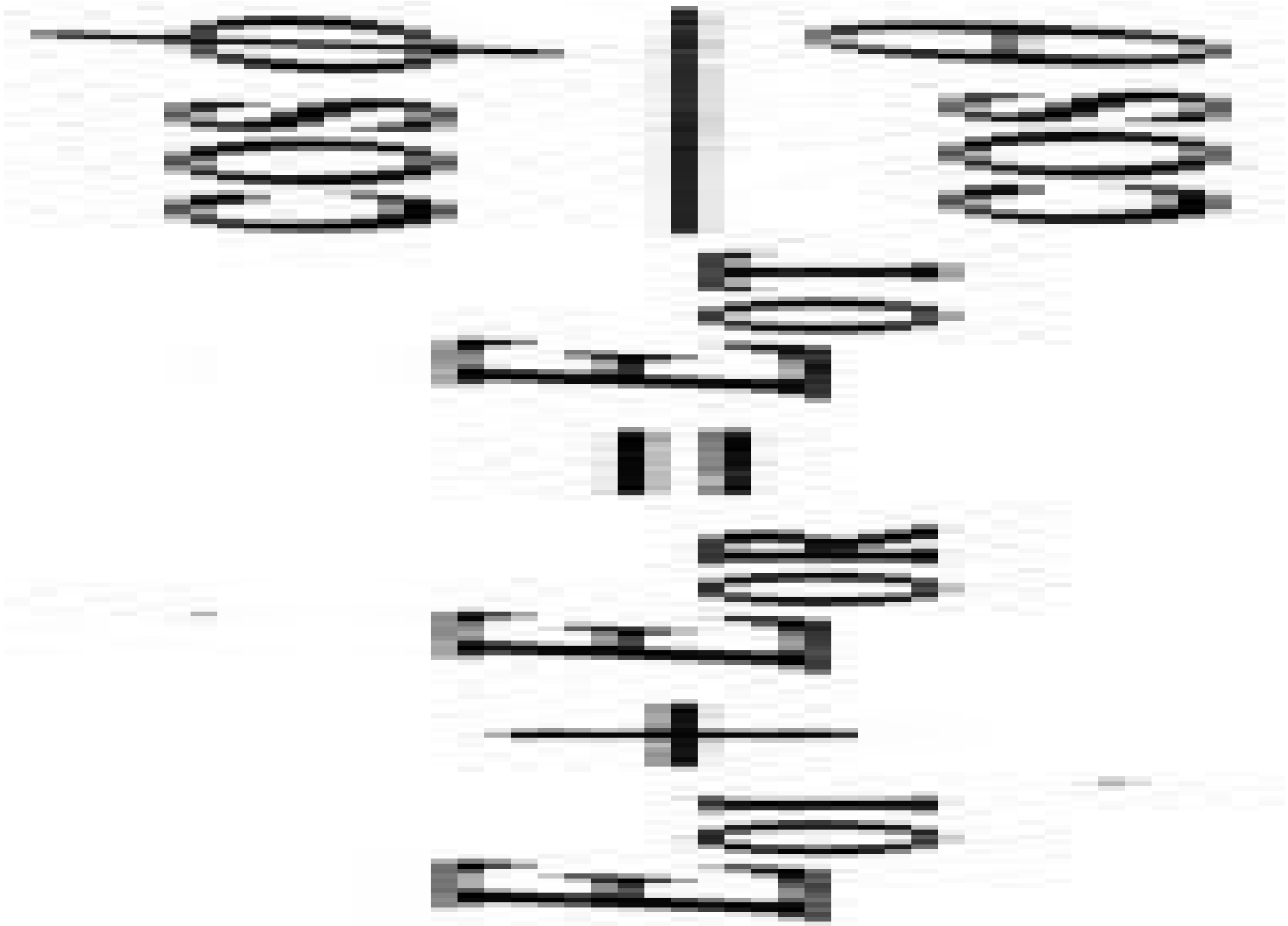
The reflection coefficient for the incident wave can be obtained by satisfying the boundary conditions on the electric fields at the point 0 in the surface, which we choose to be the origin of coordinates so that the position vector r is equal to zero. In the absence of free charge on the surface Equations (11.48) and (11.54) give, respectively, the following relationships:



and



These in turn reduce to

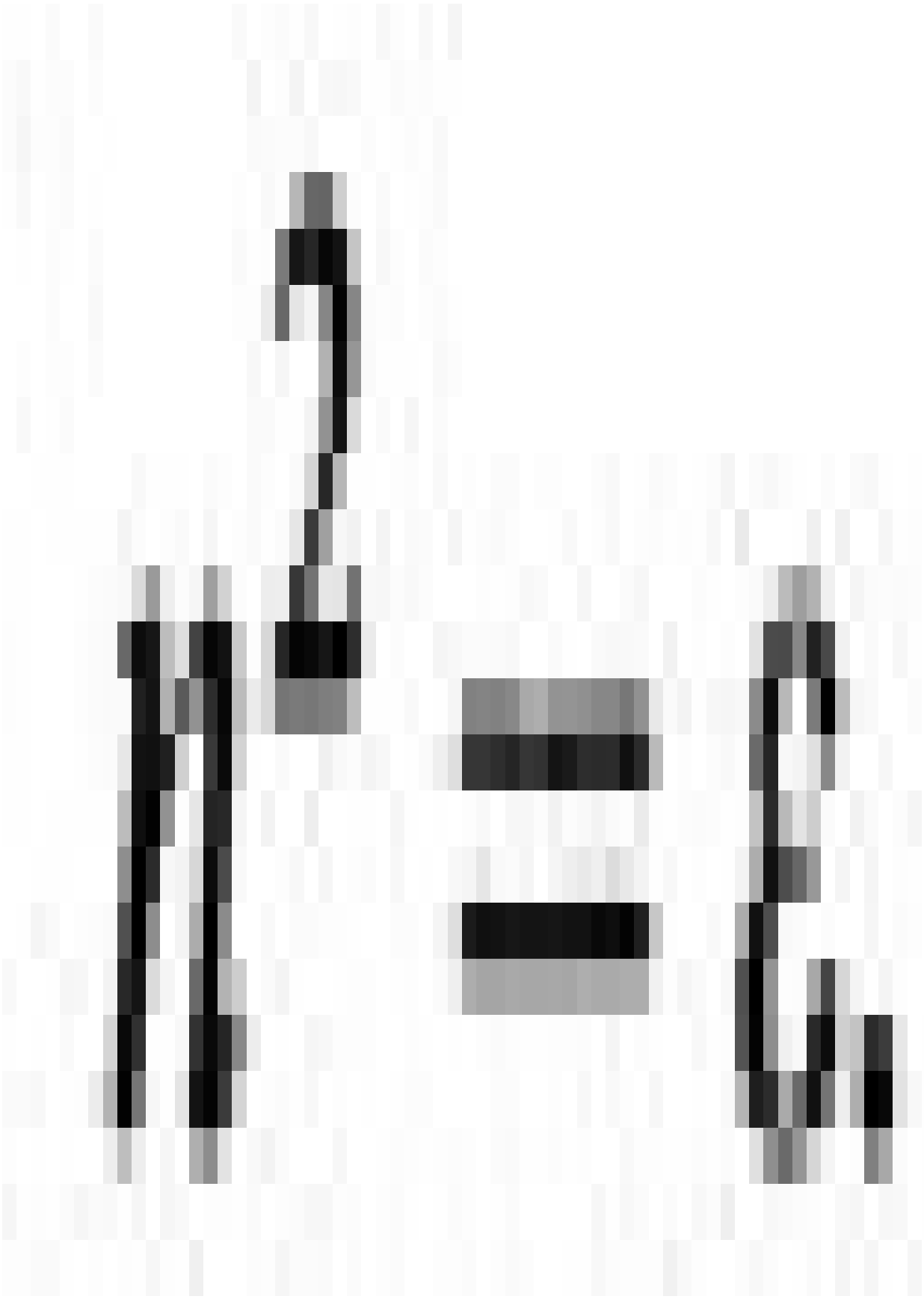


and

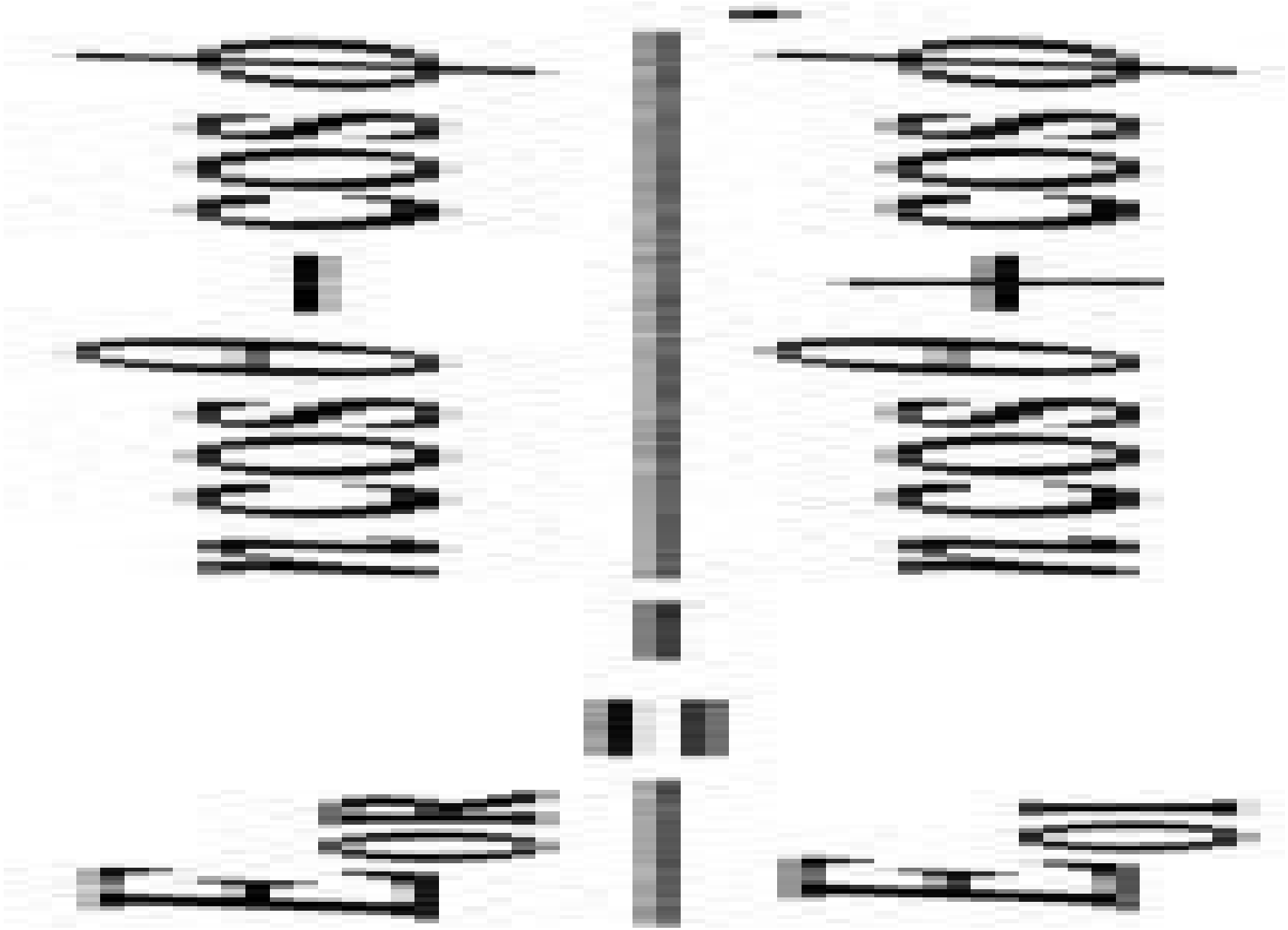
$$-E_{0I} + E_{0R} = -\varepsilon E_{0T} \frac{\sin\phi}{\sin\theta}$$

$$= -nE_{0T}$$

since



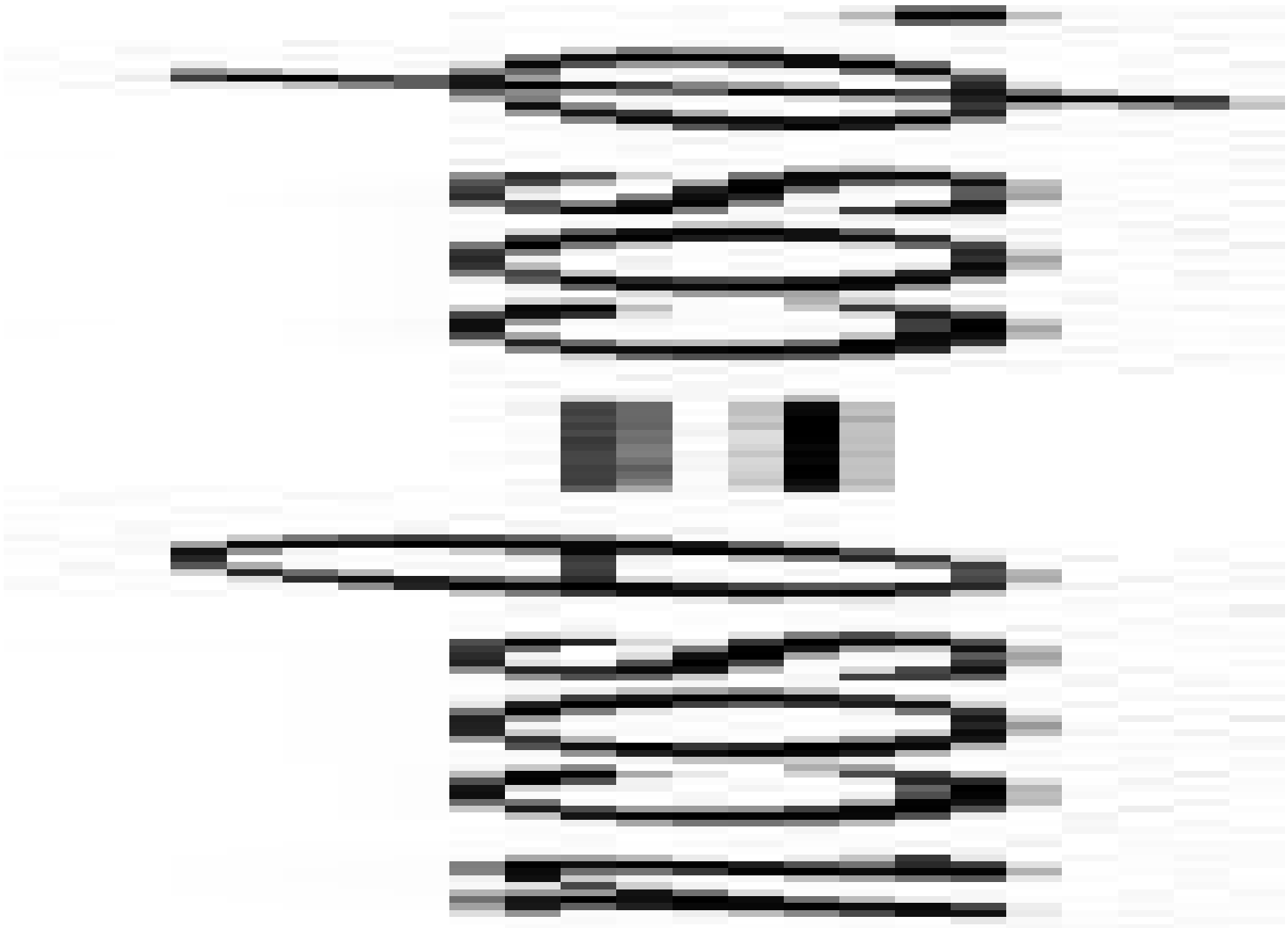
It is a straightforward matter to eliminate E0T from these two coupled equations to obtain



The reflection coefficient R is thus given by

(11.70)

It can be seen at once that there is no reflected wave whose electric vector lies in the plane of incidence if



After some algebraic manipulation this condition reduces to

(11.71)

This equation gives the Brewster angle, for which the reflected wave is zero. An analysis similar to the above for that component of the original wave with the electric vector parallel to the boundary plane shows that there is no angle of reflection for which the amplitude of the reflected wave is zero. Hence if one reflects a randomly polarized wave at a dielectric with the angle of incidence equal to the Brewster angle the result is a plane polarized wave whose electric vector lies in a direction parallel to the boundary.

For yellow light on glass, the refractive index is about 1.5, and the Brewster angle is ~ 56°. Sunlight reflected off the sea or ground is partially polarized and a great deal of the glare is cut out by using polaroid sunglasses, which will not transmit light polarized in the Earth’s horizontal plane.

Reflection at oblique incidence can also be used as an analyser of the degree of polarization of a plane wave. The small value of the reflection coefficient, given by Equation (11.70), makes this method less convenient to use than some other methods.

11.7 ELECTROMAGNETIC WAVES AND PHOTONS

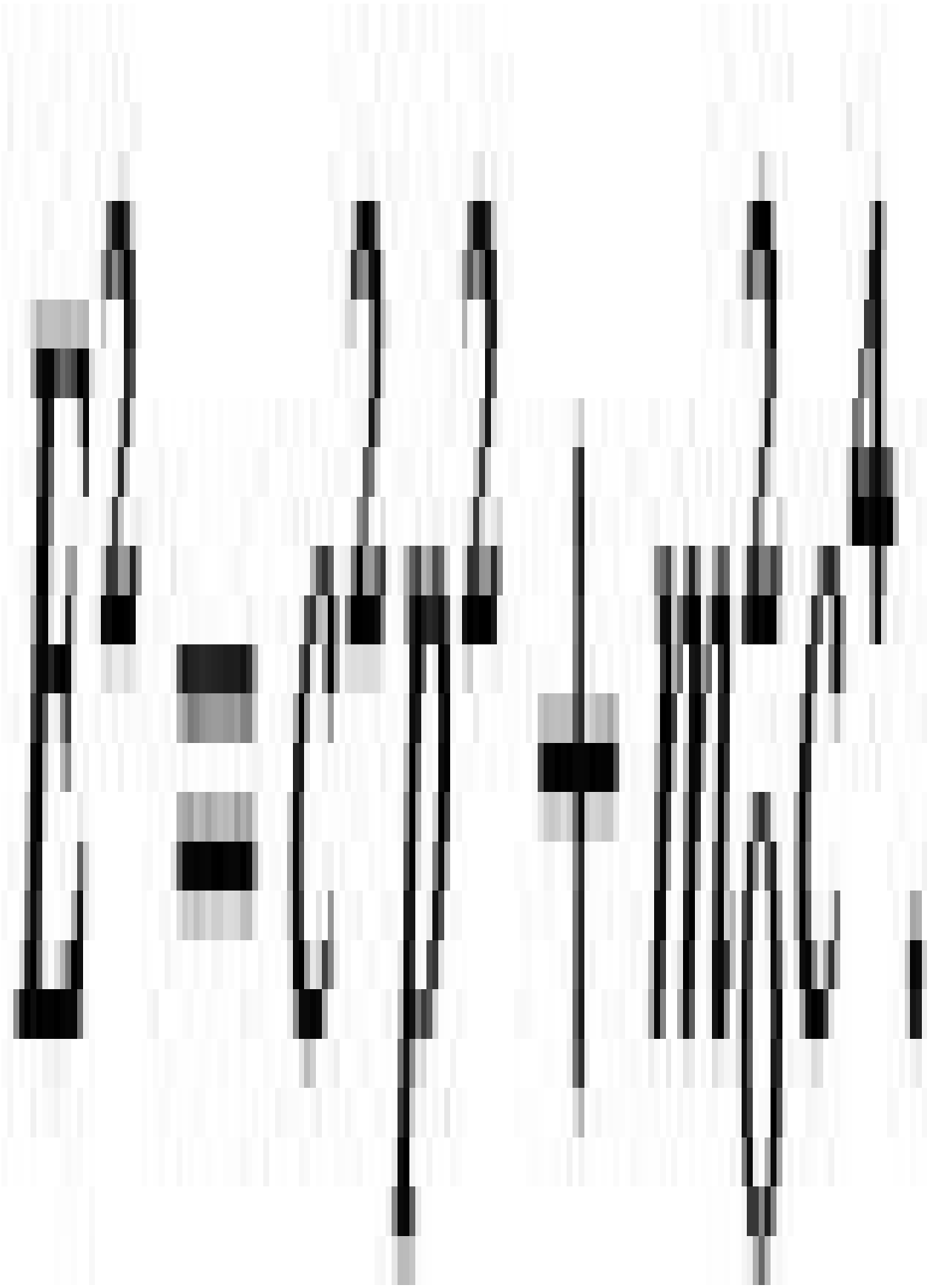
In section 11.2.2 we described classically how a wave of fixed frequency travels with a different speed according to whether it is moving in free space or in a medium. If the relative permittivity of the medium is greater than unity the wave slows down. What is the microscopic description of the slowing down process? In asking this question we have already overstepped the bounds of continuum theory, which considers the medium to be continuous. The continuum theory breaks down in any event when the wavelength of the radiation is comparable with the spacing of the molecules. This spacing is a few Ångstroms in solids. At high frequencies, and for a microscopic description of many of the phenomena discussed in this chapter, it is convenient to look upon the electromagnetic field in a different way.

The classical theory considers the energy in the field to be distributed continuously, with an energy density $(\mathbf{E}\cdot\mathbf{D} + \mathbf{B}\cdot\mathbf{H})/2$, and to flow continuously at a rate per unit area given by the Poynting vector $(\mathbf{E} \times \mathbf{H})$. Quantum theory considers the energy in an electromagnetic field to exist in discrete light quanta, called photons. Each photon moves with the speed of light in free space and has zero rest mass. The energy E of the photons depends on the frequency v of the radiation and is given by

(11.72)

where h is Planck’s constant. A simple microscopic picture of the slowing down of a wave in a block of dielectric is in terms of a cloud of photons entering the block and being exchanged back and forth between the molecules before emerging. This increases the time it takes the photons to traverse the block and corresponds to a slower speed of travel in the medium than in free space.

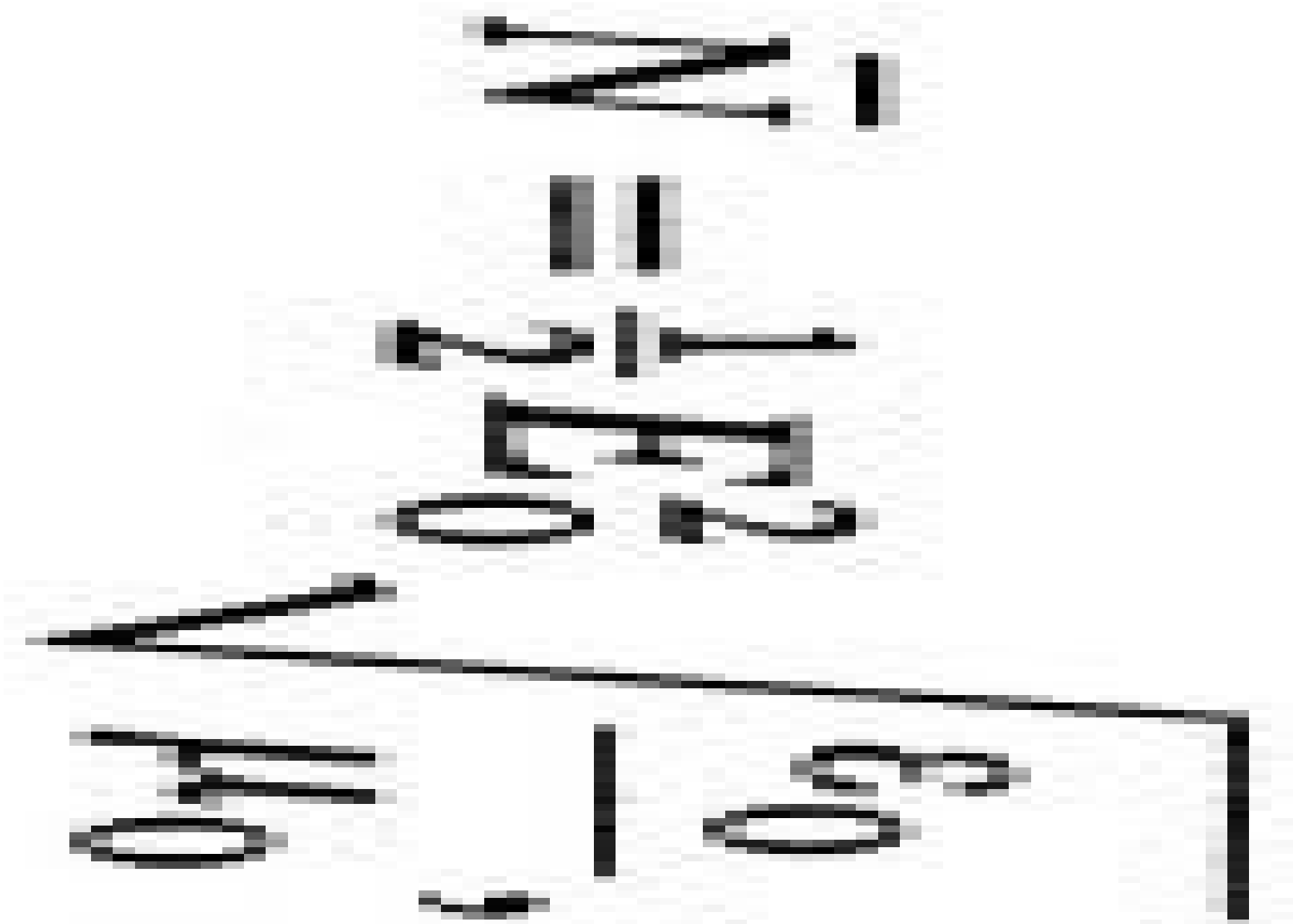
We may use the relativistic relation between the total energy E and momentum p of a body of rest mass m0 to calculate the momentum of a photon:



The rest mass of photons is zero, and the momentum of a photon of energy E is given by

(11.73)

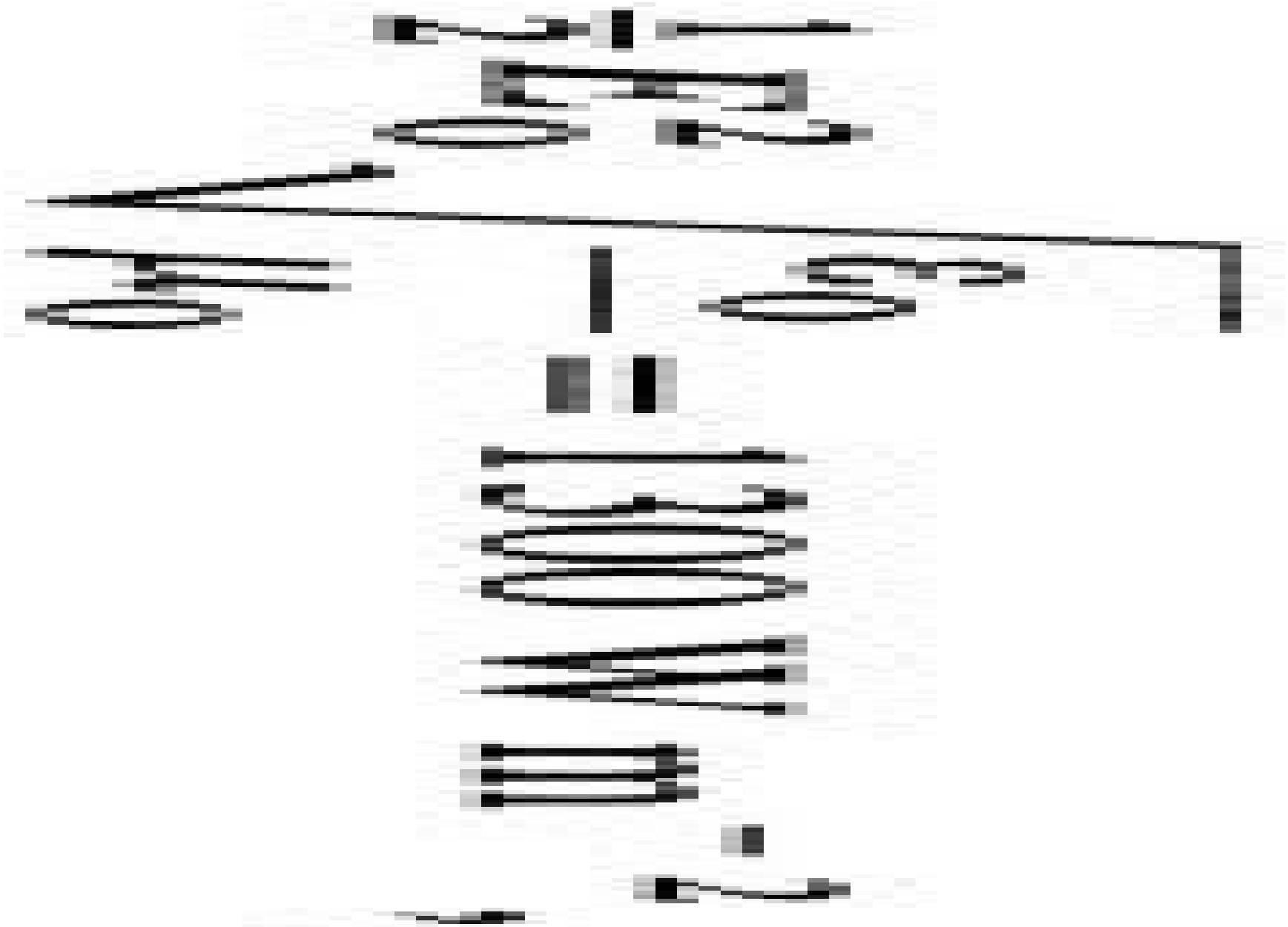
If we consider the energy in a wave to be carried by photons it can readily be seen that a plane wave incident on a metallic surface and reflected off it will exert a pressure on the surface. If the average value of the Poynting vector in the wave is , the average momentum crossing unit area per second is given by Equation (11.73) to be /c. If all the wave is reflected at a perfect conductor the total change of momentum per unit area per second is equal to 2 /c and this is the pressure P that the wave exerts on the conducting surface. If the amplitude of the electric field in the wave is E0 and the wave is incident in air or free space we may write



from Equation (11.34). Hence

(11.74)

We may make a rough estimate of the radiation pressure exerted by sunlight on the Earth if we make the assumption that all of the 1300 W m−2 falling normally on the Earth arrives in a monochromatic wave. If E0 is the amplitude of the electric field in this wave, we have



which gives a value for E_0 of $\sim 10^3$ V m^{-1}. Assuming that all of the radiation falling on the Earth is absorbed, the radiation pressure will be one half of the value given by Equation (11.74). This works out at about 10^{-5} N m^{-2}.

The radiation pressure on a conducting medium can be described classically by calculating the force on the conduction electrons (moving under the influence of the electric field of the wave) due to the magnetic field of the wave. (See Problem 11.13.) Some other electromagnetic phenomena can only be simply described if one uses the photon picture. One such phenomenon is the emission of electrons from certain materials when radiation of sufficiently short wavelength is shone on their surfaces. This effect is called the photoelectric effect. It is found that the number of photoelectrons emitted increases with the intensity of the wave shone on the surface of the material, but that the maximum energy of the emitted electrons is constant for a fixed frequency. Quantum mechanically an increase in the intensity of the radiation corresponds to an increase in the number of photons in the wave, the energy of the photons being constant at fixed frequency. Thus from the standpoint of the light quantum hypothesis both the results on the maximum energy of the photoelectrons and on their number can be readily understood.

PROBLEMS 11

11.1 The root mean square of the displacement current density in a linearly-polarized monochromatic plane wave in free space is 10^{-5} A m^{-2}. The frequency is 10^8 Hz. Obtain values for the amplitudes of the electric and magnetic fields in the wave.

11.2 Show that the time average of the energy density in a monochromatic linearly-polarized plane wave moving in an isotropic non-conducting medium is distributed equally between the electric and magnetic fields.

Show that in a conducting medium the average energy in the magnetic field is greater than in the electric field.

11.3 A laser is a device that emits a parallel beam of monochromatic light. The intensity may be assumed constant across the beam. If the power is 1 watt and the beam has a diameter of 1 mm calculate the maximum amplitude of the magnetic field B in the beam in free space.

11.4 The centres of two close circular metal plates forming a parallel plate capacitor, initially charged, are connected internally by a fine wire of negligible self-inductance. Obtain an expression for the Poynting vector at the surface of the wire at any subsequent instant and hence evaluate the total flow of energy into the wire.

11.5 Deduce the conductivity of a medium of relative permittivity equal to five if the magnitude of the conduction and displacement current densities in it are equal when a monochromatic plane wave of frequency 10^8 Hz is propagated.

What is the attenuation per metre of the amplitude of a plane wave of frequency 10^6 Hz when propagated in such a medium? The relative permeability is unity.

11.6 A capacitor, formed by two flat parallel plates of area 113.1 cm^2, spaced 1 mm apart in air, is connected in series with a coil, wound with a single copper wire 40 m long and 0.2 cm in diameter, and an alternating voltage generator. If the inductance of the coil at resonance is 1 mH what is the Q of the circuit if the resistivity of copper is 1.72 μΩ cm?

11.7 The skin depth of 1 MHz radio waves in pure copper is approximately 10^{-4} m at room temperature. Calculate the skin depth of 1 MHz radio waves at 15 K when the conductivity has increased by a factor of 10^4.

11.8 The electrical conductivity of sea water is about 4 (Ω m)−l. What is the skin depth for low frequency radio waves of wavelength 3000 m?

11.9 At heights greater than about 50 km above the surface of the Earth, in the ionosphere, the gas molecules in the atmosphere are ionized by ultra-violet rays from the Sun. If the density of free electrons is n0 per unit volume, determine the relative permittivity of the ionosphere for waves of frequency v. (There are magnetic fields in the ionosphere due to the motions of the charges, but these can be neglected when calculating the relative permittivity.)

11.10 A monochromatic plane wave in free space is incident normally on the plane surface of a medium of refractive index equal to two. If the amplitude of the electric field in the incident wave is 10 V m−1, what is its value inside the medium?

11.11 Explain why the relative permittivity of water vapour begins to decrease at a lower frequency than the frequency at which that of nitrogen gas begins to decrease.

11.12 Show that if a material has a relative permeability greater than unity, at a certain angle of incidence there will be no reflection of a plane wave incident with its electric field vector parallel to the boundary.

11.13 Calculate classically the radiation pressure exerted on a conducting surface by a perpendicularly incident plane electromagnetic wave, and in the limit of infinite conductivity show that the answer reduces to Equation (11.74).

11.14 The intersteller medium is a dilute plasma containing free electrons and ions with density 3×10^4 m−3. A pulsar which is at a distance of 10^{19} m from Earth emits a short pulse of electromagnetic radiation containing frequency components covering the spectrum from visible light down to radiofrequencies. Find the difference in arrival times at Earth between radiation corresponding to red light and radio-frequency radiation at 100 MHz.

* Looking in the direction of propagation of the wave, i.e. looking in the + z-direction, the electric field vector, at a fixed point, rotates anticlockwise as time progresses. The wave (11.15) is said to be a left handed circularly polarized wave. Note that the E vector rotates clockwise as z increases if the wave is observed at a fixed time, looking in the direction of propagation.

* For a discussion of wave and group velocities see, for example, Vibrations and Waves, by A.P. French, Van Nostrand, London, 1982.

CHAPTER 12

Waveguides

The voltage and current waves along a coaxial cable or a pair of parallel metal plates were discussed in Chapter 9. The starting point of the discussion was that of electrical circuits, proceeding from the time delays introduced by successive sections of a ladder network to the smooth propagation of signals on transmission lines. It turned out that the velocity of propagation is always equal to the velocity of electromagnetic waves in the medium in which the line is embedded.

Now that we have considered the boundary conditions that electromagnetic waves have to obey at the surface of conductors, we can think of transmission lines in a different way, regarding them as systems that guide electromagnetic waves along certain paths. The conclusions of Chapter 9 could have been obtained by fitting solutions of Maxwell's equations to the boundary conditions appropriate to the conductors of the line.

A transmission line of a given shape may not be a very efficient way of transmitting power or signals from one point to another. For example two parallel wires radiate energy and there are heating losses in the wires: a coaxial cable with the outer cylinder earthed produces no fields outside, but there remain the energy losses due to resistive heating of the conductors, chiefly the inner one where the fields are largest. The most efficient way of transmitting energy over short distances at frequencies greater than about 1000 MHz is by using a waveguide. This is a hollow metal conductor whose cross-section is of a simple and constant shape. The walls reflect the waves to and fro, and there exists the possibility of combining the fields in the reflected waves in such a way as to add up to a wave that is propagated down the length of the pipe with very little

attenuation. If the metallic walls were perfectly conducting there would be no attenuation at all since all of the incident energy would be reflected every time. This would be so if the metal used for the surfaces of the guide walls were superconducting. It is possible to made waveguides with superconducting walls; however, the costs associated with their use are prohibitive. For waveguides normally used there are small energy losses at each reflection due to the currents induced in the walls by the transmitted fields. The energy losses in a length δx of the guide at a distance x from its beginning are proportional to the length δx and to the energy in the guide at that point. The mean energy in the wave in a waveguide thus falls off exponentially with distance travelled. In contrast, when radiation is emitted from antennas (this topic is discussed in the next chapter), the energy crossing unit area per second in the radiation fields falls off inversely as the square of the distance from the antenna. Hence if one is comparing antennas and waveguides from the point of view of energy losses it is usually best to use antennas to transmit waves over long distances. This is because at large distances the exponential rate of energy decay in a waveguide always results in a smaller energy than the inverse square law fall off in free space. Waveguides, however, are much better over shorter distances and are widely used in practice for the transmission of microwaves of wavelengths in the range from a few millimetres to about one metre. For example, 3 cm radar waves are usually generated in an area near the deck of a ship where they can be controlled, and are carried to the antenna on the top of the ship’s mast with waveguides.

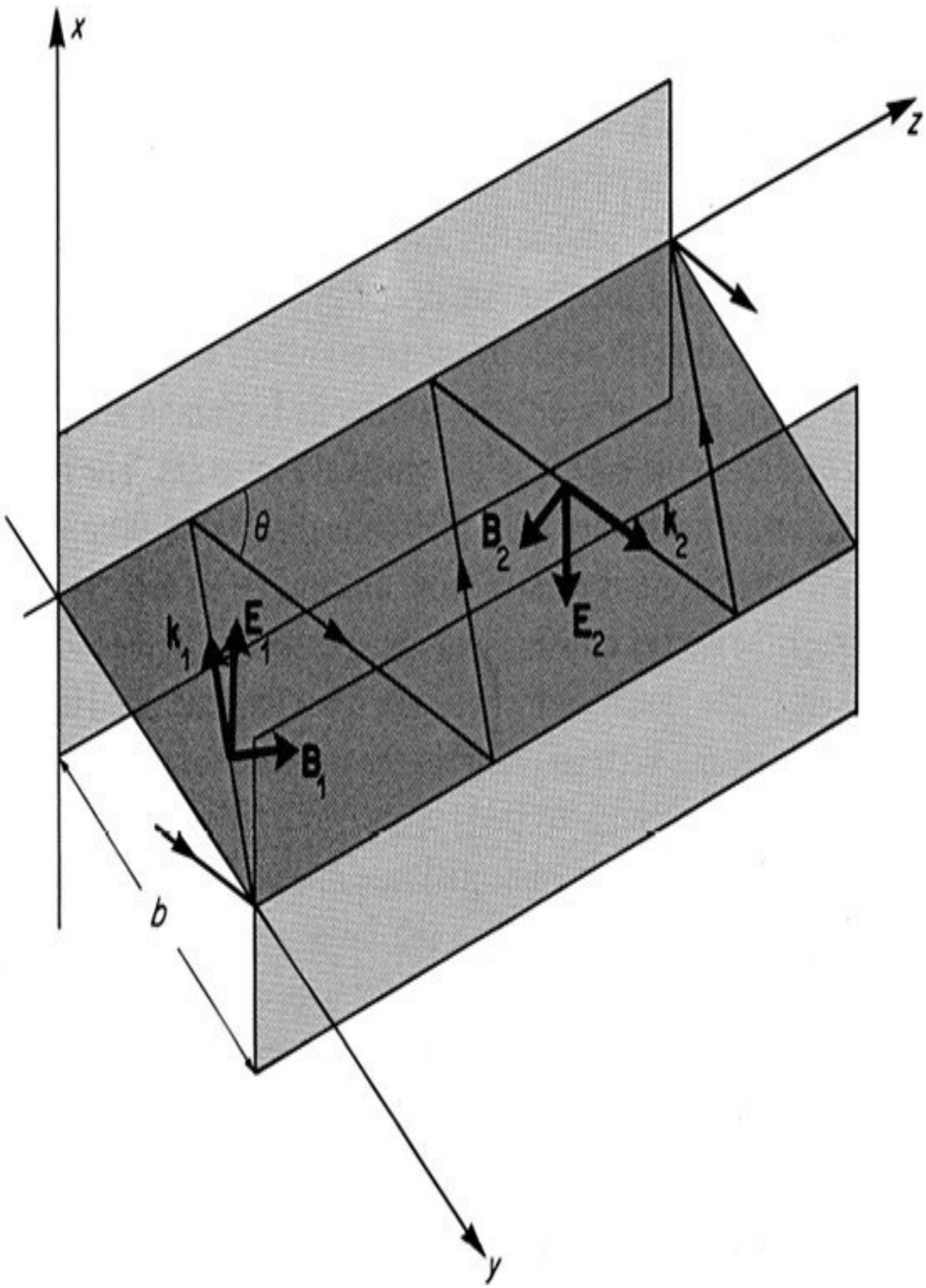
Waveguides of rectangular cross-section are the most commonly used and we shall, for the most part, discuss these alone.

12.1 THE PROPAGATION OF WAVES BETWEEN CONDUCTING PLATES

The main features of the propagation of waves in waveguides can be understood by considering the simpler problem of the propagation of waves between two parallel conducting plates.

Figure 12.1 shows parts of two parallel plates separated by a distance b. We will suppose that the plates are of infinite extent and furthermore that they are perfectly conducting. This last assumption is made in order to simplify the mathematics. The problem now is to determine the characteristics of the electromagnetic fields which can be propagated down the length of the plates. The fields we are looking for have to obey Maxwell's equations in the free space between the plates. They also have to obey certain boundary conditions at all points on the conducting walls. These conditions, as discussed in section 11.6.1, are that the tangential component of the electric field and the normal component of the magnetic field must be everywhere zero over the walls. The normal component of the electric field need not be zero, since there can be charges on the conducting surfaces, and the tangential component of the magnetic field need not be zero, since there can be surface currents in the perfectly conducting walls. The condition that the fields obey the last two of Maxwell's equations in the space between the plates means, just as in section 11.1, that they must satisfy the wave Equations (11.5) and (11.6). The simplest type of solution to such wave equations is a plane wave, and it is immediately clear that the linearly polarized plane wave in which the fields are

Figure 12.1. Parts of two parallel, infinite, perfectly conducting plates. One plate is in the plane y = 0, the other in the plane y = b. A wave propagating in the z-direction between the plates is made up of a sum of waves travelling in the directions of the vectors k1 and k2.


x
z
y
θ
k1
E1
B1
B2
E2
k2
b

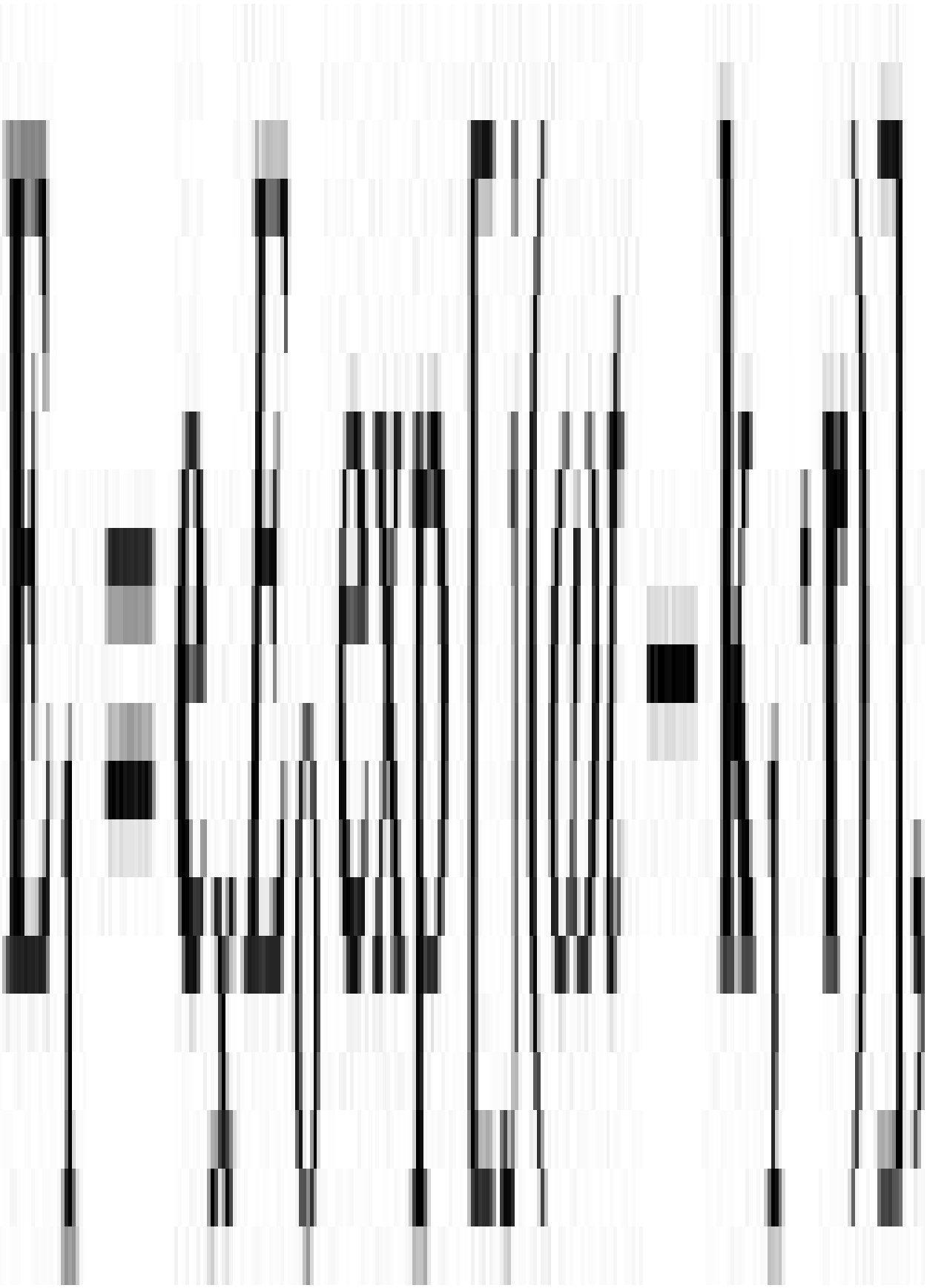
(12.1)

and

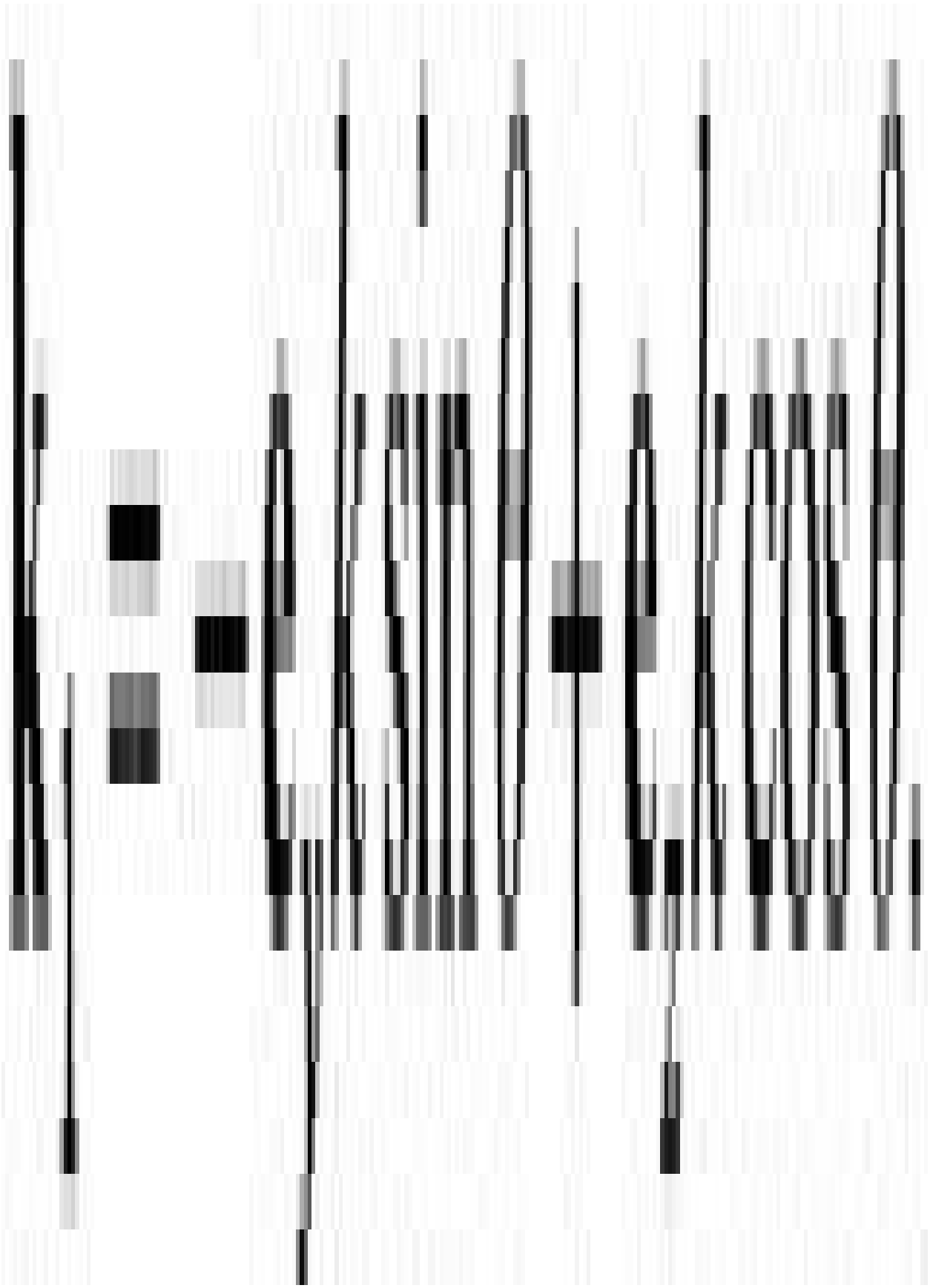
(12.2)

satisfies all the boundary conditions at the walls. Waves of this form at any frequency can thus be propagated down the plates. The wave is called a transverse electric and magnetic or TEM wave because both the electric and magnetic fields are transverse to the direction of propagation. TEM waves, different to the plane wave above*, can be propagated along any pair of conductors of constant shapes, and they travel with the speed of light in the medium in which the conductors are embedded. These are the waves travelling on transmission lines which were discussed from a different point of view in Chapter 9. Inside a hollow, single, conductor like a waveguide, however, TEM waves cannot be propagated. We will now try to find more general types of wave which satisfy the boundary conditions at the parallel plates: this will lead us to a solution of the waveguide problem.

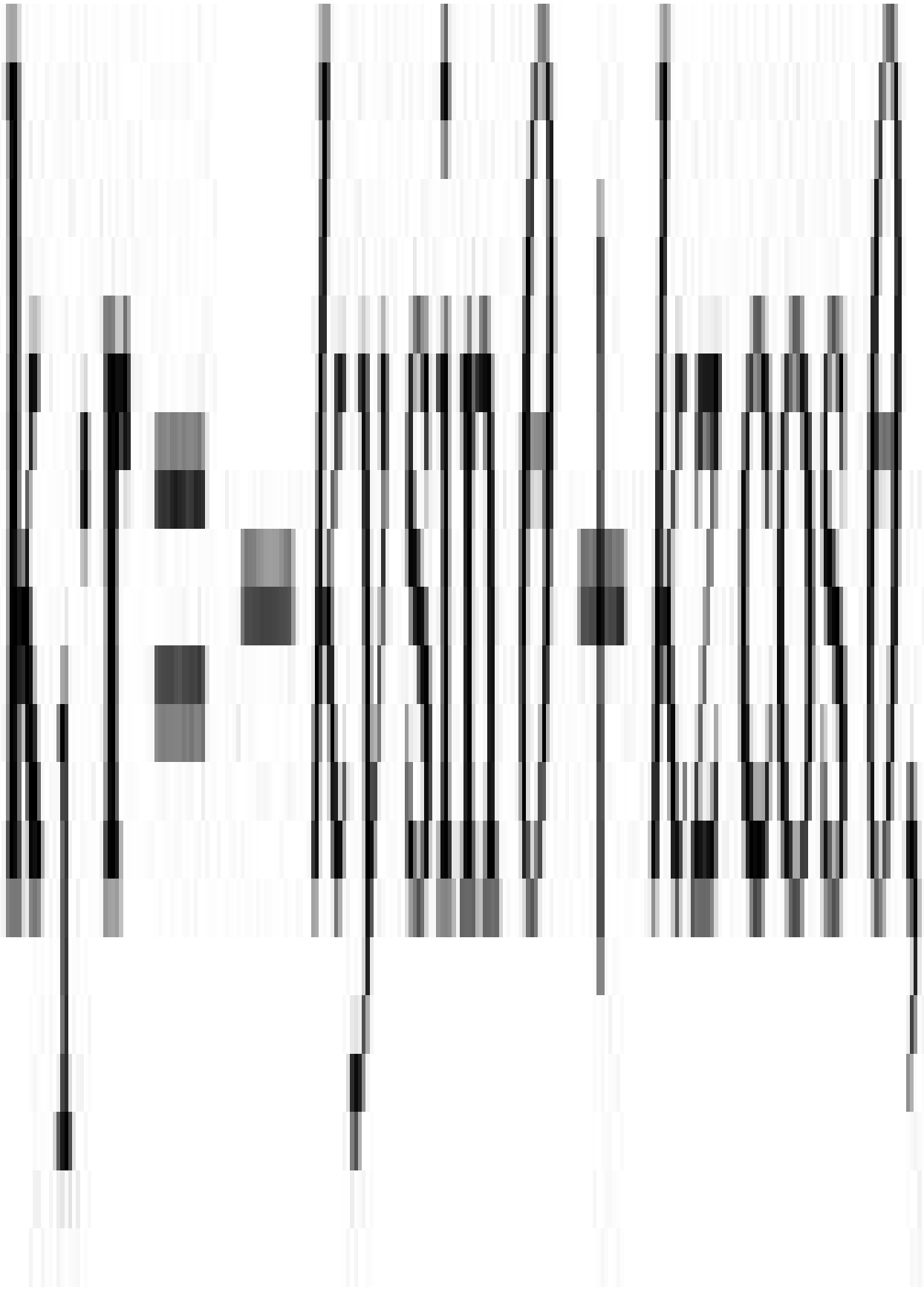
Any superposition of plane waves between the plates is a solution to the wave equations, and we will try to find a combination of waves which satisfies the boundary conditions on the fields at the walls. Let us choose axes of coordinates as shown in Figure 12.1 and suppose that a linearly polarized plane wave is moving between the plates in a direction given by the vector k1 This vector lies perpendicular to the x-axis and makes an angle θ with the z-axis as shown. We will take the electric field vector in the wave to oscillate along the x-direction with amplitude E0. The expression for the electric field in the plane wave is given by Equation (11.18)



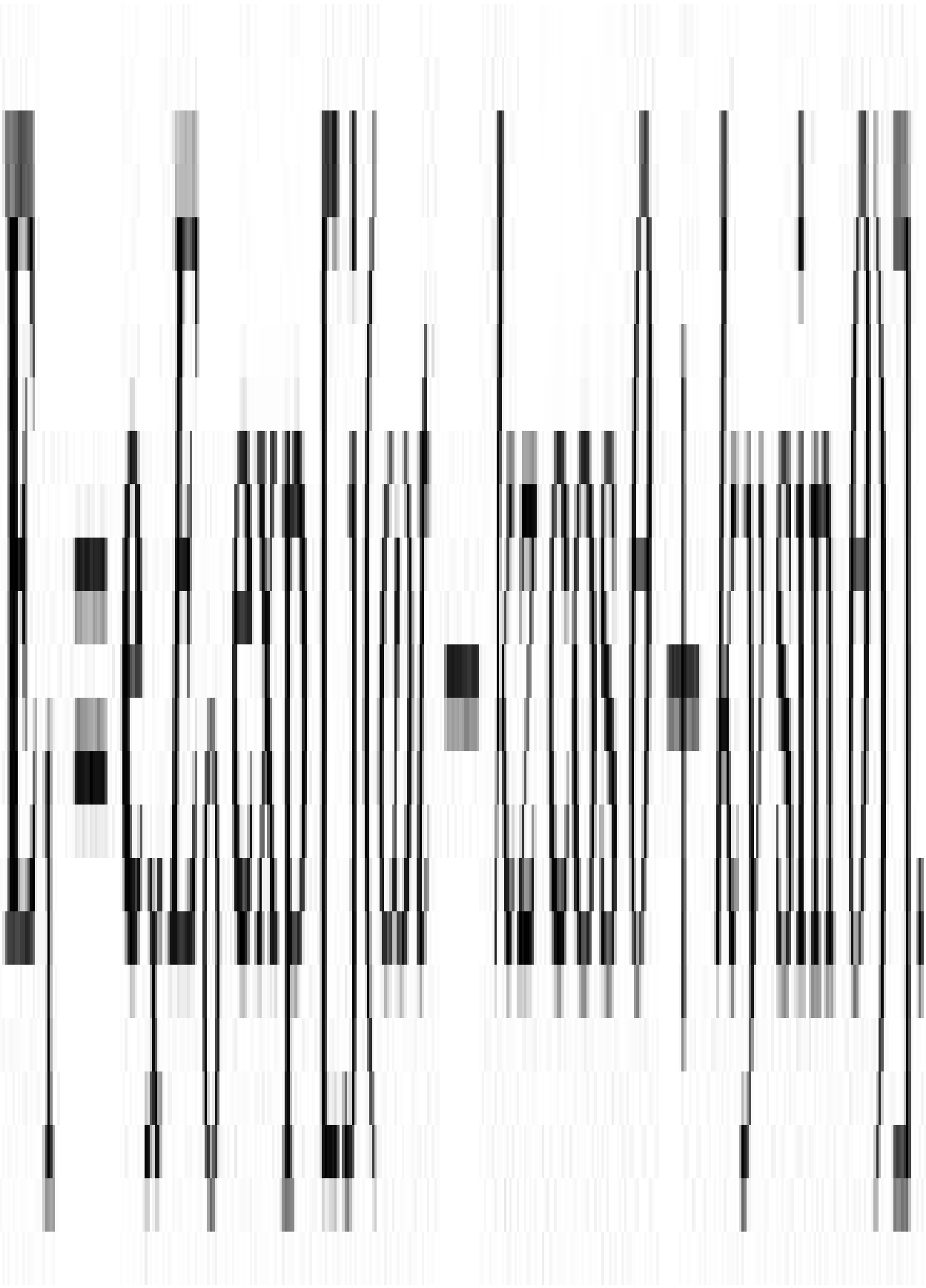
where ω and k, the modulus of k1, are the angular frequency and wavenumber of the wave in free space. The vector k1 is given by



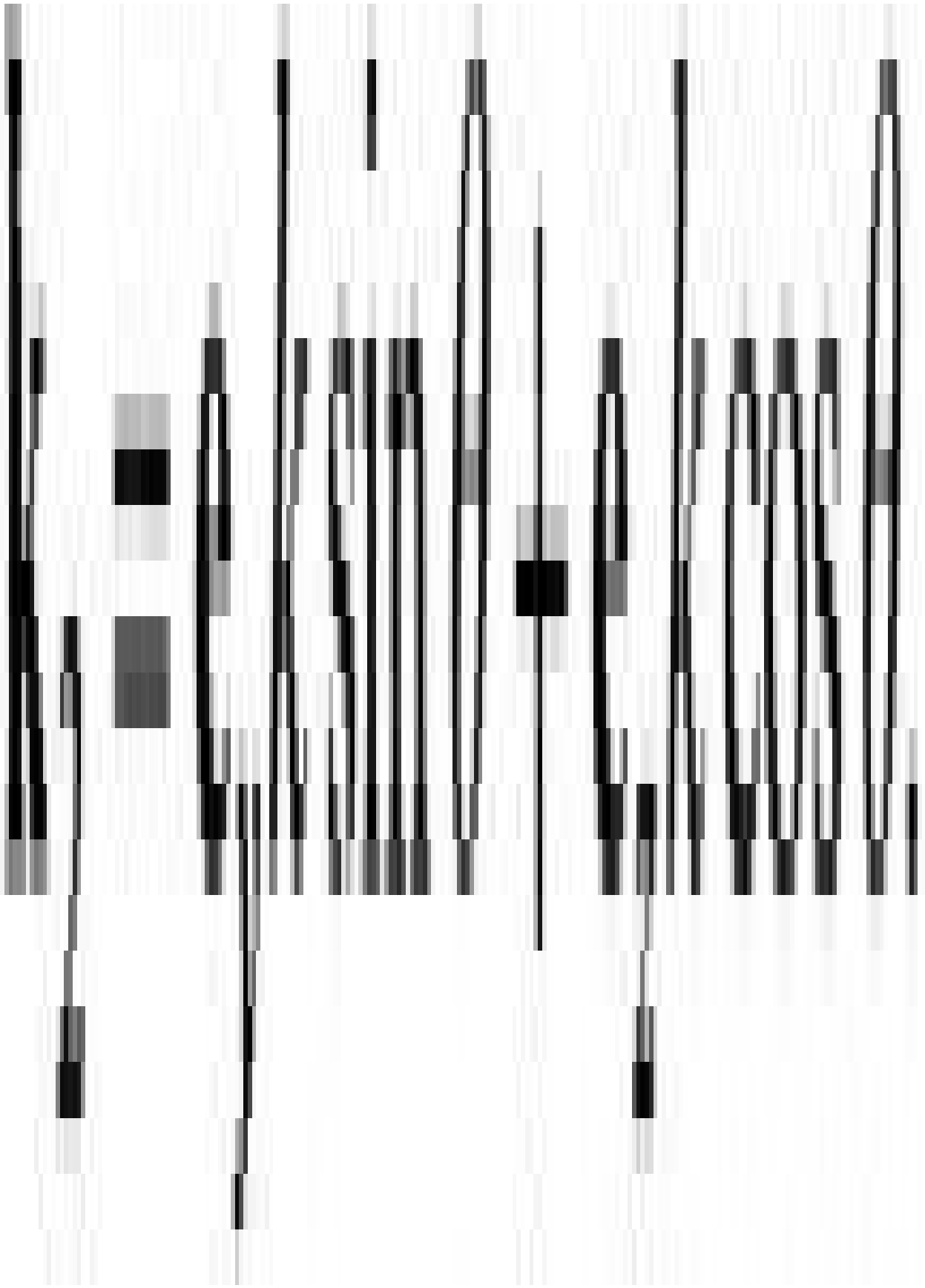
Hence



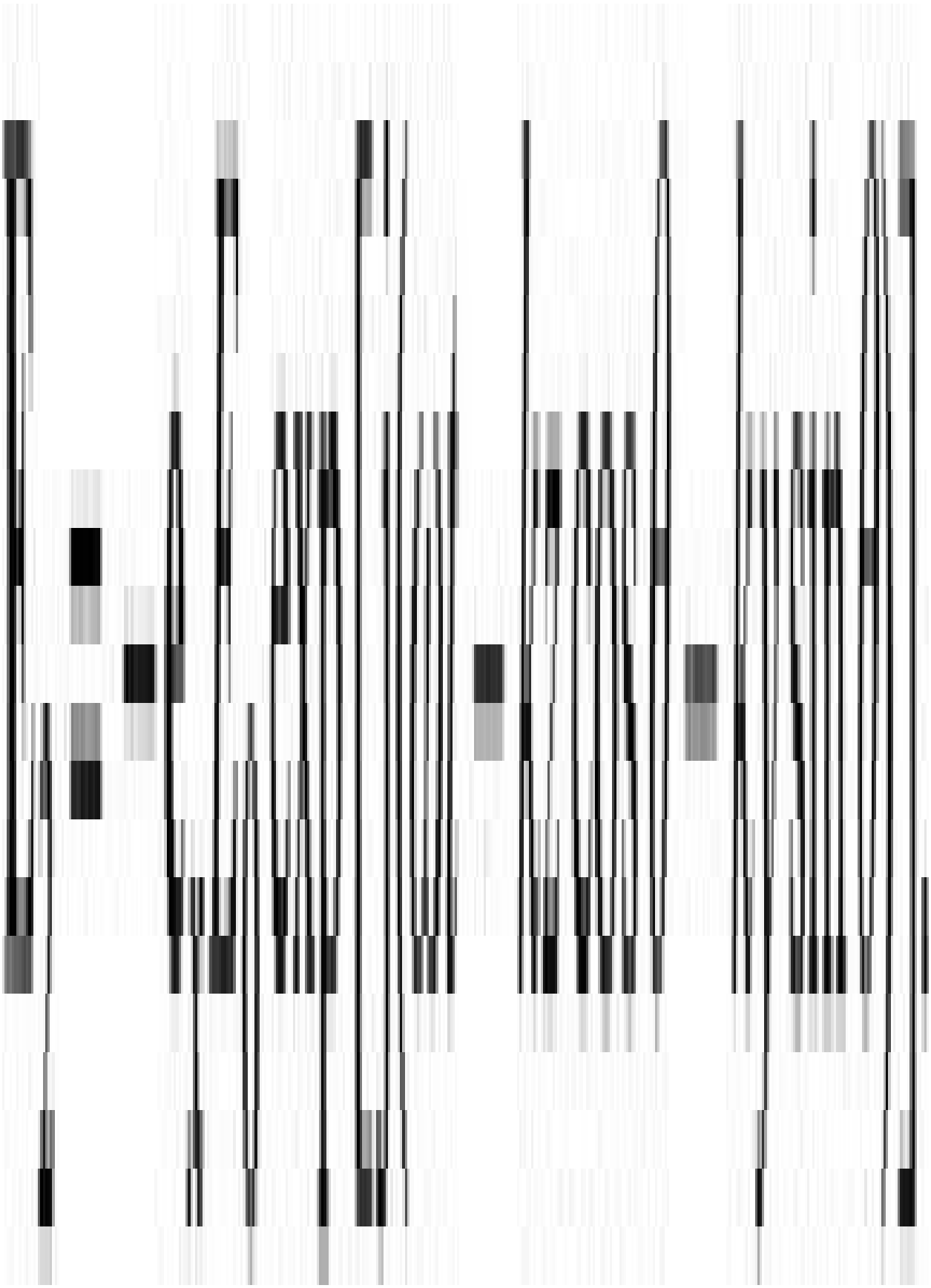
and the wave may be written



When this wave strikes the plane y = 0 it is totally reflected, with a phase change of 180°, in the direction of the vector k2, given by



The magnitude of the vector k2 is also equal to k, and the electric field in the reflected wave is



Let us now see if the superposition of two fields described by the functions E1 and E2 can be made to fit the boundary conditions. The resultant field E is given by

(12.3)

$$E = e_y E_0 \exp[j(\omega t - kz\cos\theta)]$$

$$\times [\exp(jky\sin\theta) - \exp(-jky\sin\theta)]$$

$$= e_y 2jE_0 \sin(ky\sin\theta)\exp[j(\omega t - kz\cos\theta)].$$

The boundary condition that Ex = 0 at y = 0 is automatically satisfied. We can make the tangential field zero over the other plane by choosing the angle θ such that Ex = 0 at y = b, i.e. such that

(12.4)

where n is an integer. There may be several different acceptable fields, corresponding to waves propagating in the z-direction, depending on the different possible values of the integer n. Since sin θ cannot exceed unity there is a limitation on the number of different fields. For a given frequency of the radiation the maximum value of n is given by the condition

(12.5)

If $\pi/kb \geq 1$, i.e. $\lambda/2b \geq 1$, no wave at all of the type we have been considering is possible. There is thus a cut-off frequency below which such waves cannot be propagated. This cut-off frequency v1 can be obtained by putting n equal to unity and taking the equality in the expression (12.5):

(12.6)

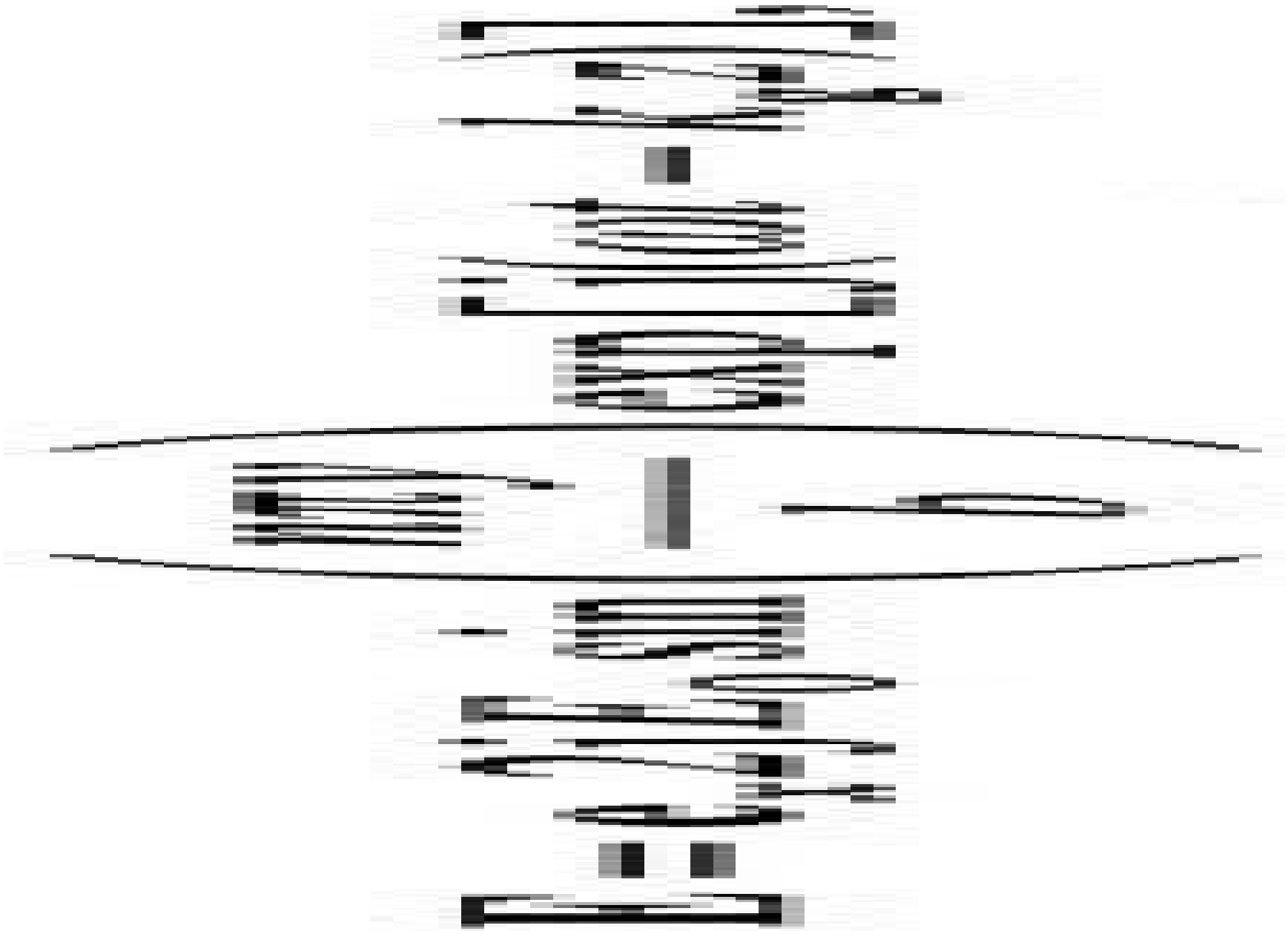
The different waves that correspond to the different allowed values of the integer n are called modes. Each different mode has its own cut-off frequency vn which can be determined from Equation (12.5) to be

(12.7)

The cut-off frequencies of each mode increase as the order of the mode increases, i.e. as the value of the integer n increases. For a plate separation of 1.5 cm, radiation of free space wavelength 2 cm (frequency 1.5×10^{10} Hz) can only be propagated in the lowest, n = 1, mode.

Substitution of the expressions for sin θ and cos θ obtained from Equation (12.4) into Equation (12.3) gives the electric fields in the waves in terms of the integer n.

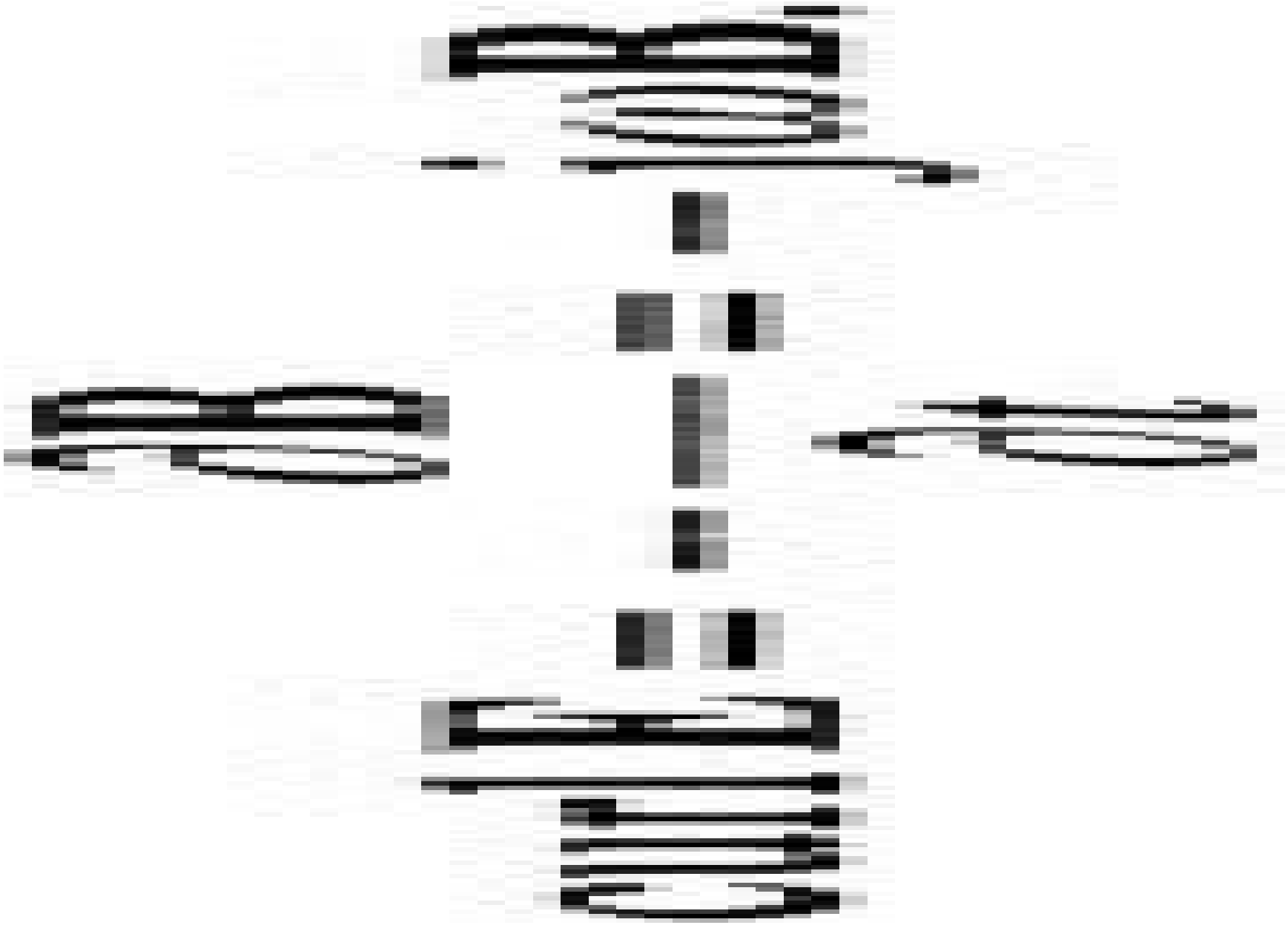
(12.8)



where

(12.9)

kg is equal to $2\pi/\lambda g$ where λg is the wavelength of the wave motion in the z-direction. The electric fields in the waves described by Equation (12.8) have no components in the z-direction, the direction of propagation. The waves are thus called transverse electric waves, and the wave of order n in designated TEn. The magnetic fields in the TEn waves can easily be obtained from the relation



The expressions for the different components of the electric and magnetic fields are given below:

(12.10)

$$E_x = 2jE_0 \sin\left(\frac{n\pi y}{b}\right) \exp\left[j(\omega t - k_g z)\right]$$

$$E_y = E_z = 0$$

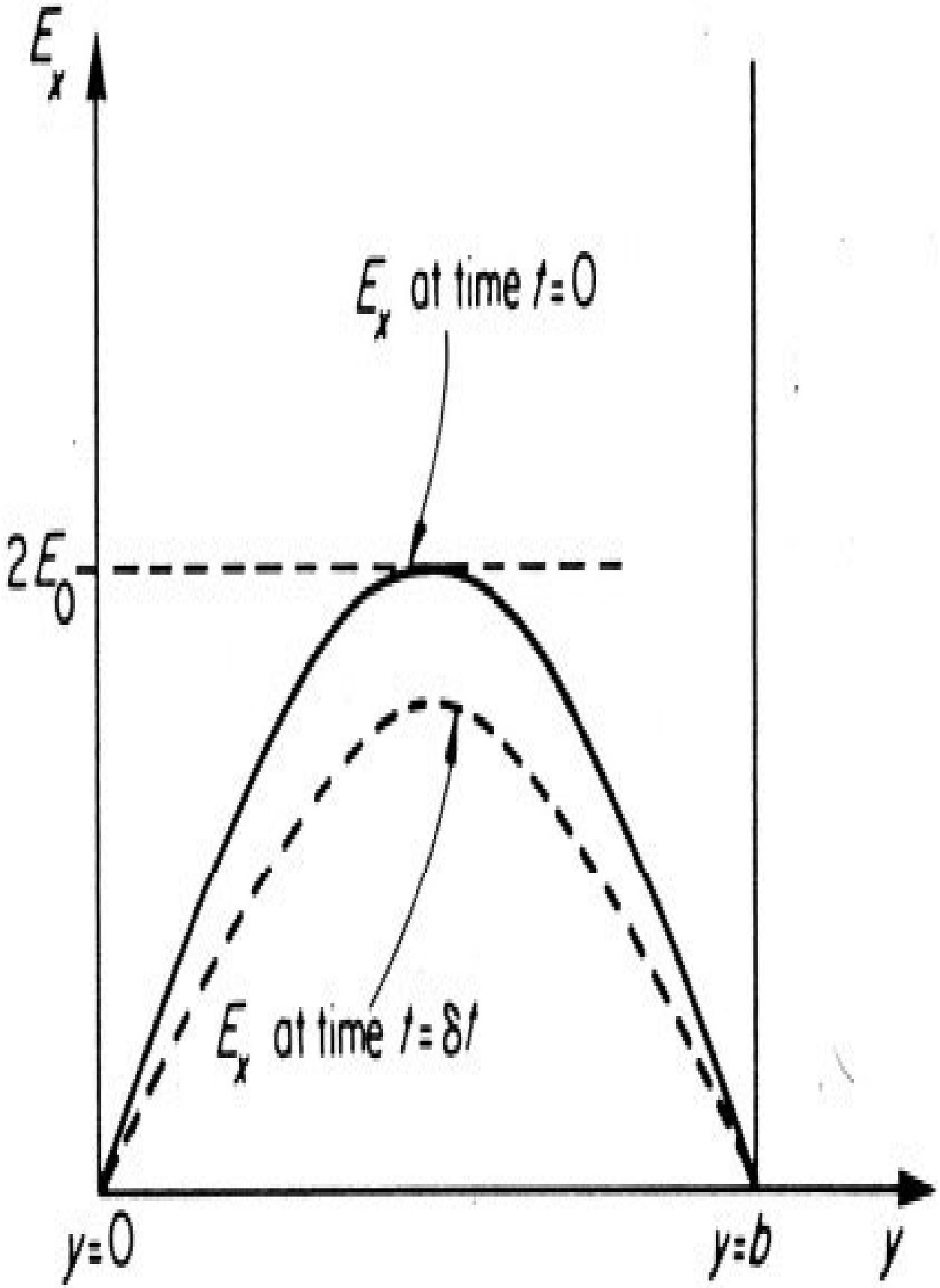
$$B_x = 0$$

$$B_y = 2jE_0 \frac{k_g}{\omega} \sin\left(\frac{n\pi y}{b}\right) \exp\left[j(\omega t - k_g z)\right]$$

$$B_z = 2E_0 \frac{(k^2 - k_g^2)^{1/2}}{\omega} \cos\left(\frac{n\pi y}{b}\right) \exp\left[j(\omega t - k_g z)\right].$$

We may make several observations on these expressions. (i) They are independent of the coordinate x, as required by the symmetry of the problem. (ii) The fields are made up of a superposition of plane waves and therefore satisfy Maxwell's equations. (iii) The component By satisfies the boundary condition that the normal component of B be zero at the walls, i.e. at y = 0, and y = b. (iv) The wave propagates in the z-direction, and so the Poynting vector must be in this direction. Since Ex is the only non-zero component of the electric field, the field B must have a y-component in phase with Ex. The field B has a z-component $\pi/2$ out of phase with Ex so that the energy moving in the y-direction averages to zero.

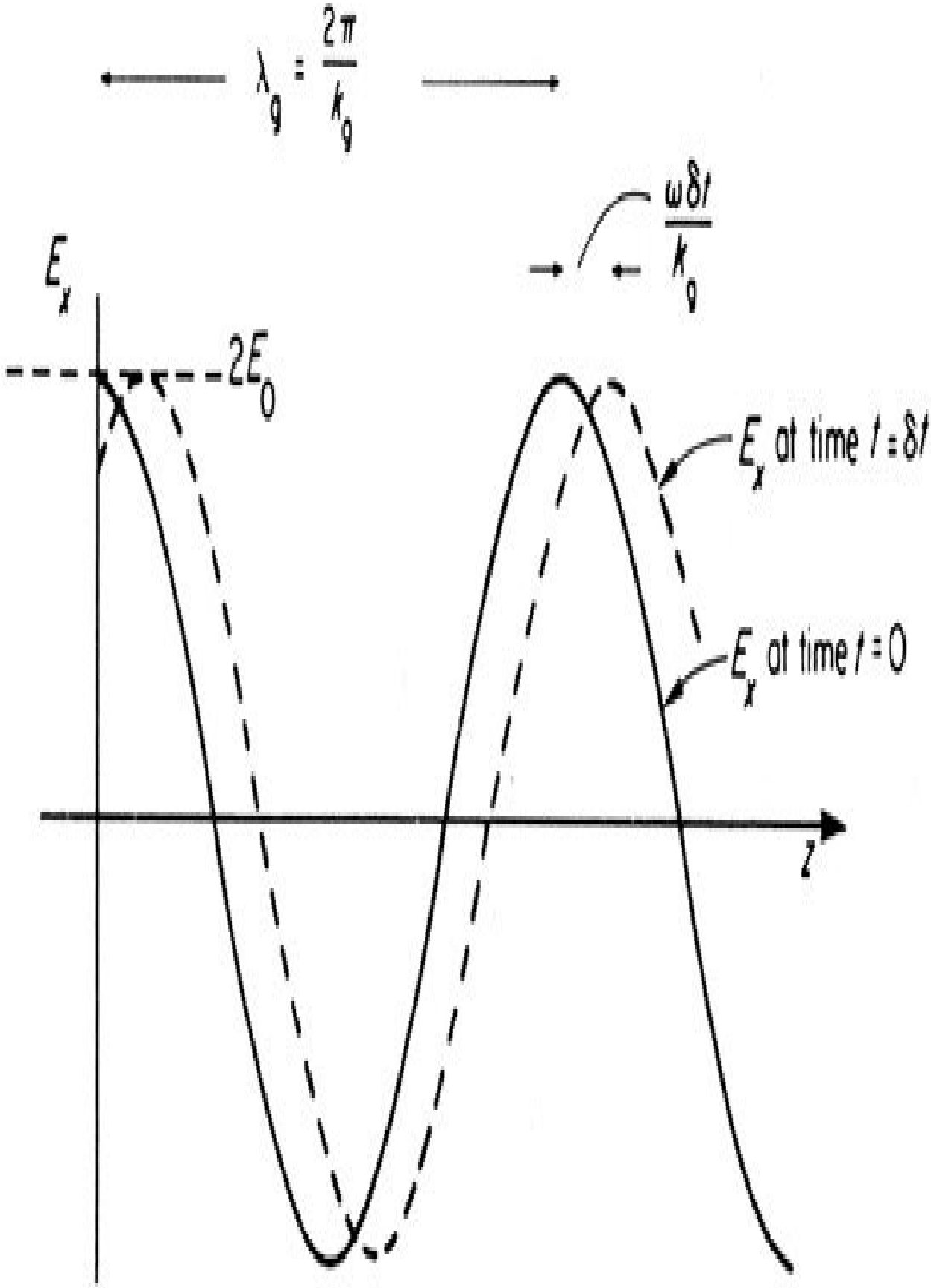
Figure 12.2. The electric field at position z = 0 and time i = 0 in the TE1 mode propagating between parallel plates. The field is independent of the coordinate x, and at any point oscillates in time with angular frequency ω.


E_x
E_x at time $t = 0$
$2E_0$
E_x at time $t = \delta t$
$y = 0$
$y = b$
y

The variation of the electric field in the TE1 mode with the coordinate y at time t = 0 and position z = 0 is shown on Figure 12.2. The field at any point (, y, 0) oscillates with time at the angular frequency ω, with the maximum amplitude of the oscillations halfway between the plates at y = b/2. The variation of the electric field in the z-direction at time t = 0 halfway between the plates is shown on Figure 12.3. As time progresses the wave is propagated in the +z direction with wave velocity

(12.11)

Figure 12.3. The variation of the electric field in the TE1 wave with distance z at time t = 0. The field drawn is that halfway between the plates where y = b/2. The electric field in the TE01 mode in a rectangular waveguide is similar to the fields shown in this figure and on Figure 12.2


$\lambda_g = \frac{2\pi}{k_g}$
$\frac{\omega \delta t}{k_g}$
E_x
$2E_0$
E_x at time $t = \delta t$
E_x at time $t = 0$
z

Substituting into this the value of kg from Equation (12.9) we obtain

(12.12)

From this equation it can be seen that the wave velocity exceeds the speed of light in free space. However, the energy of a group of waves travelling between the planes moves with the group velocity vg, which is always less than the speed of light. vg is given by Equation (11.28)

(12.13)

This can be worked out from Equation (12.9) to give

(12.14)

The product of the wave velocity and the group velocity is thus equal to the square of the speed of light.

(12.15)

The expressions (12.10), describing acceptable waves which propagate between the plates, are obtained by adding two waves in both of which the electric field vector is perpendicular to the z-direction. The electric field in the resultant wave

is thus perpendicular to the z-axis, whilst the magnetic field does have a component in the z-direction. We can obtain another set of solutions which satisfy the boundary conditions and the wave equation and correspond to waves travelling in the z-direction by adding two waves in which the magnetic field has an x-component only. The resulting waves have no component of the magnetic field in the direction of propagation and are called transverse magnetic waves, or TM waves. Like the TE waves they are characterized by integers n which give the variation of the magnetic field in the y-direction. The different TMn modes have cut-off frequencies at the same values as the TEn modes, given by Equation (12.7). The TE and TM modes, and the TEM wave discussed earlier, are the only solutions to the wave equation that obey the boundary conditions and correspond to waves travelling in the z-direction. For radiation of a given angular frequency the most general wave propagating in the z-direction is a mixture of all the different possible modes. The relative amounts and phases of the different modes depend on the precise nature of the source of the radiation between the plates.

12.2 RECTANGULAR WAVEGUIDES

We can now consider a rectangular waveguide, which we make by adding two sides to the conducting planes of the last section. This gives us a hollow metal pipe of rectangular cross-section, with width a in the x-direction, and width b in the x-direction, as shown on Figure 12.4. The direction of propagation of the waves is the z-direction. We will continue to suppose that the metal has infinite electrical conductivity, and will determine the characteristics of the waves that can travel down the waveguide.

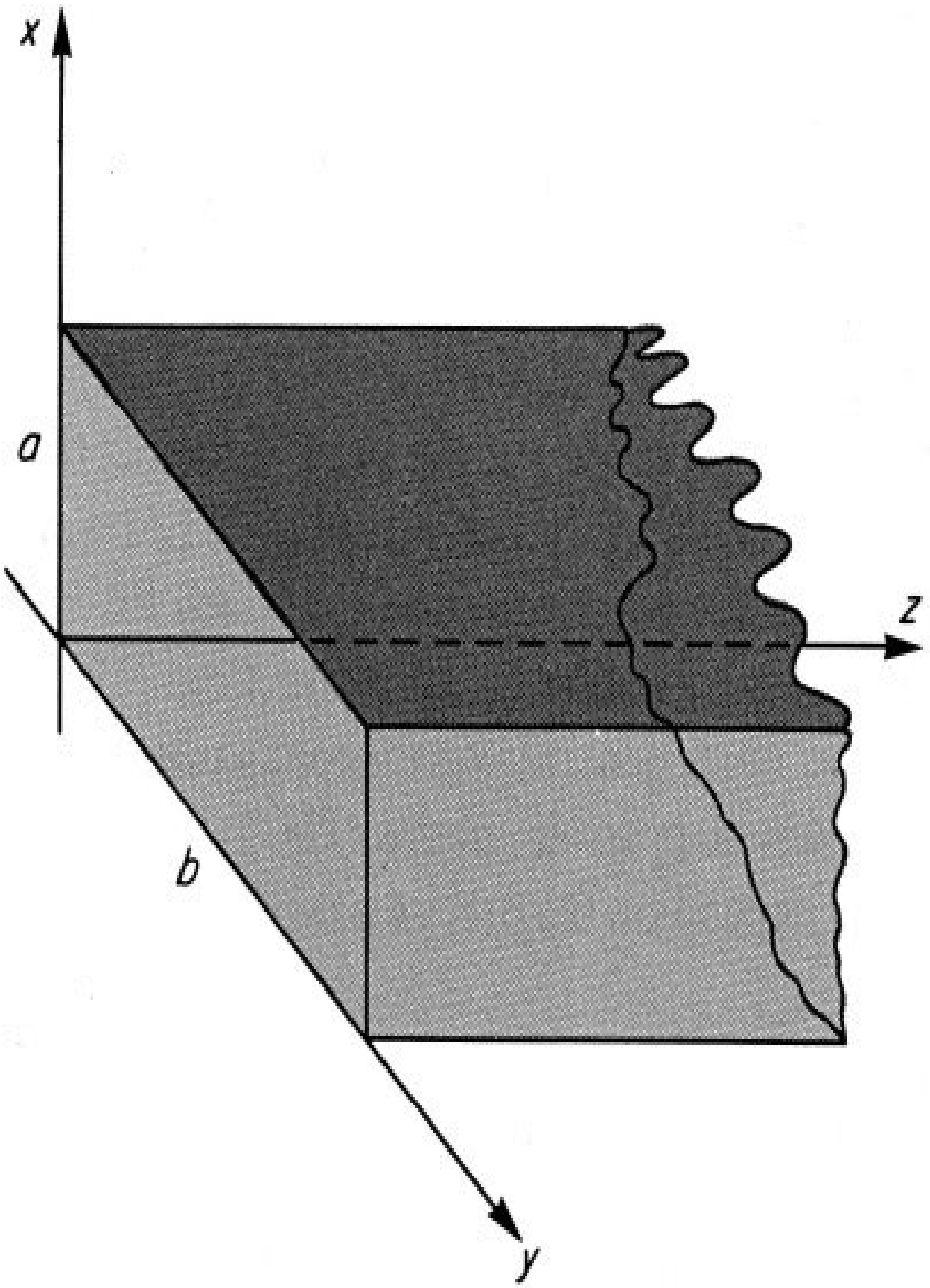
There are two possible sets of waves which can propagate without attenuation. In one set, called TE waves, the electric field is perpendicular to the direction of propagation; in the other set, called TM waves, the magnetic field is perpendicular to the direction of propagation. The TEM plane wave which was a possible wave between the parallel plates no longer satisfies the boundary conditions over the new walls, and so is not an acceptable wave within the waveguide.

We will restrict our attention at first to the transverse electric TE waves. It can be seen at once that the TEn waves of the infinite parallel plate problem are still solutions to Maxwell's equations which satisfy all the boundary conditions at the guide walls. They do not violate the boundary conditions on the new walls parallel to the y–z plane since the electric fields in the TEn waves are everywhere perpendicular to the new walls and appropriate charges are induced on the walls; the magnetic fields are everywhere parallel to the new walls and appropriate surface currents are induced in them.

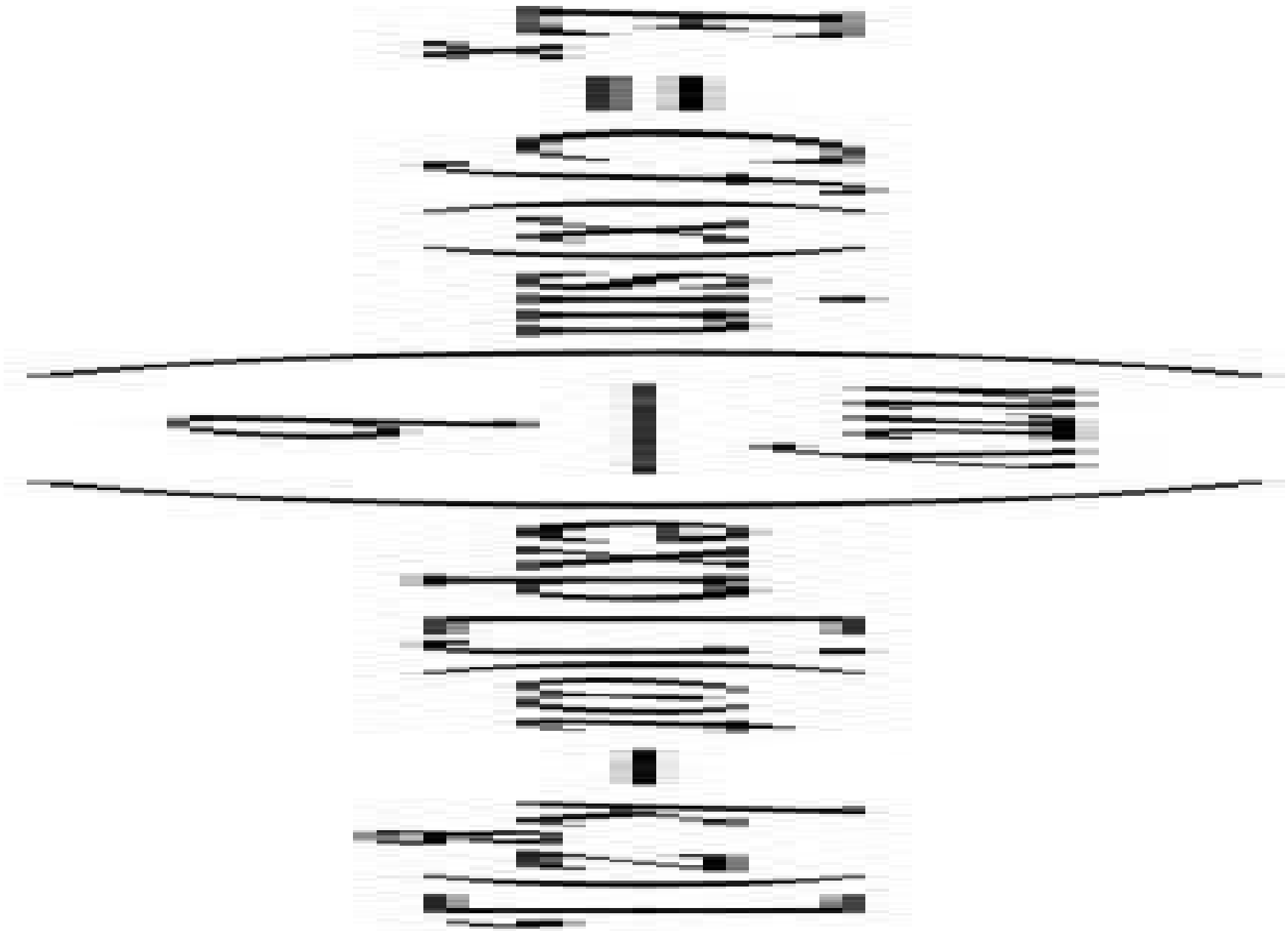
For all the TEn waves the electric fields inside the guide are independent of the coordinate x. Now that the guide is not infinite in the x-direction we may expect new solutions of the wave equation in which the field does vary in the x-

direction. These new modes are designated TEmn, the integer index m playing a similar role for the variation with x as the index n does for the variation with y. Thus we relabel the TEn modes of the parallel plate problem as TE0n modes in the wave guides. From our experience with the parallel plates we may guess that for a TEmn mode, the x-component of the electric field will be of the form

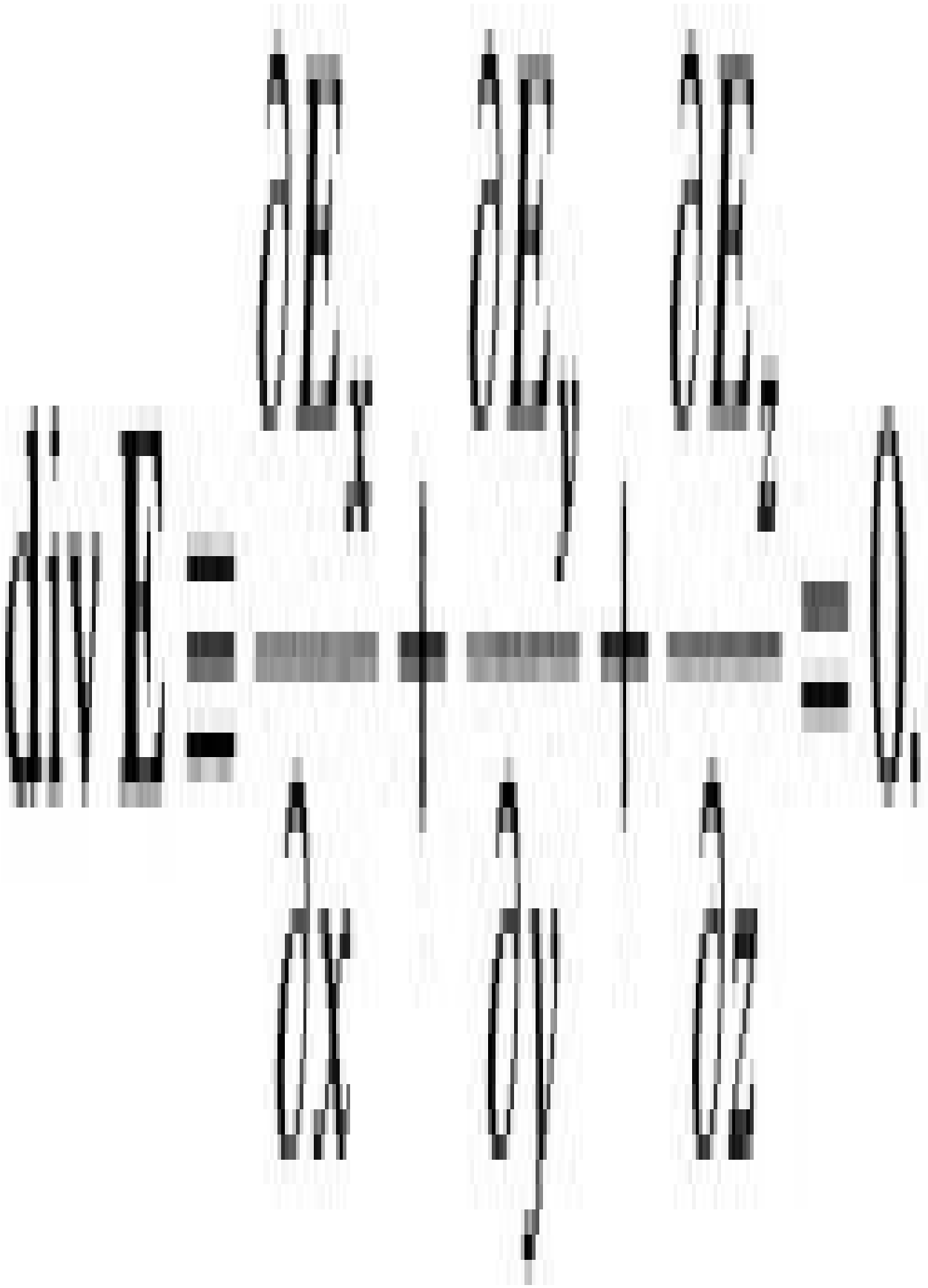
Figure 12.4. The coordinate system used in the discussion of rectangular waveguides.


x
a
z
b
y

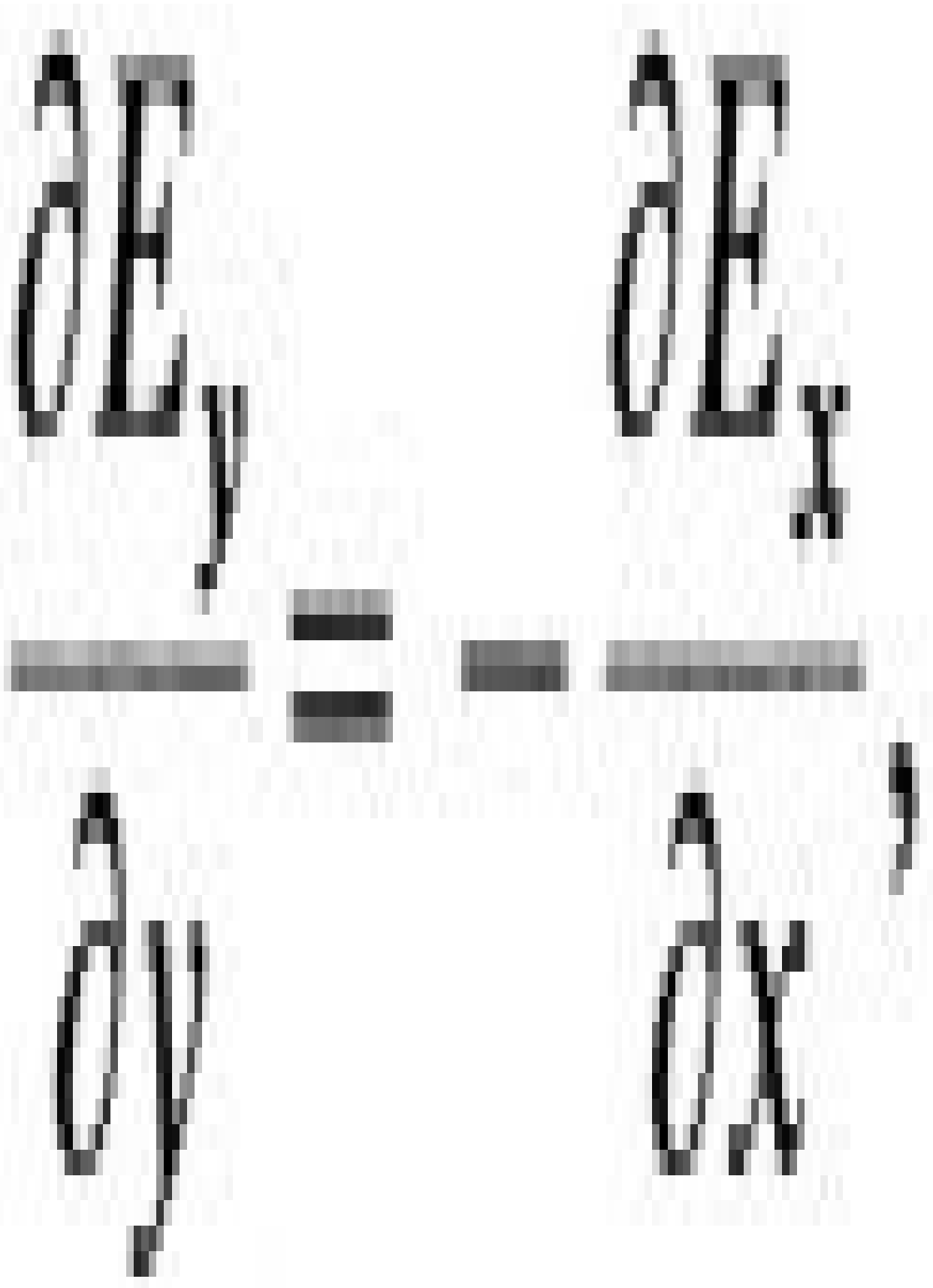
(12.16)



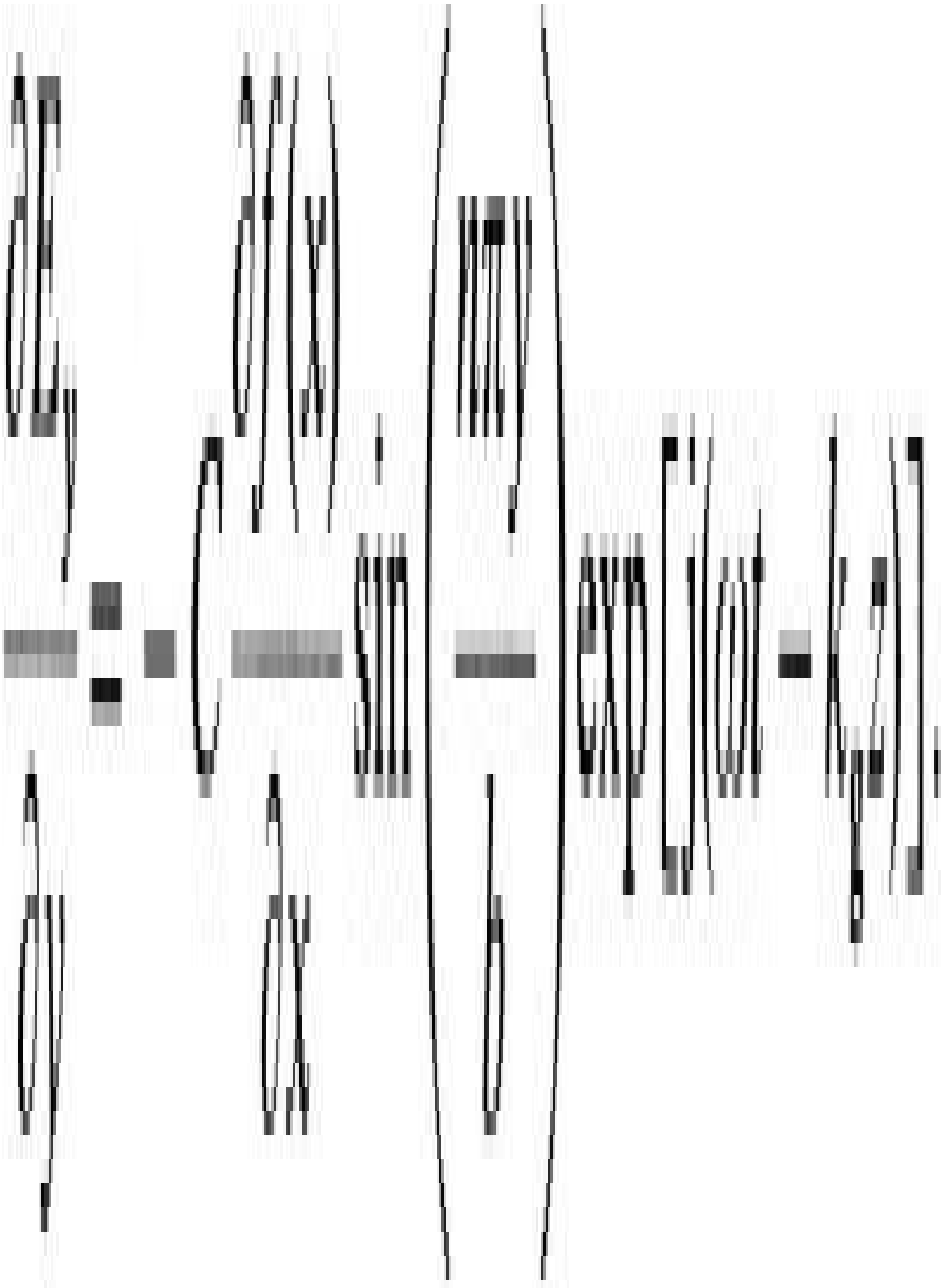
where C is a constant, and the function f(x) gives the variation of the field component in the x-direction. kg is again the wavenumber of the wave in the guide. If this expression is indeed an acceptable wave it must satisfy the wave Equation (11.6), satisfy the boundary conditions at the walls, and also be the x-component of an electric field which satisfies the Maxwell equation



Since we are seeking a TE solution, the function $\partial Ez/\partial z$ is zero by definition. Hence

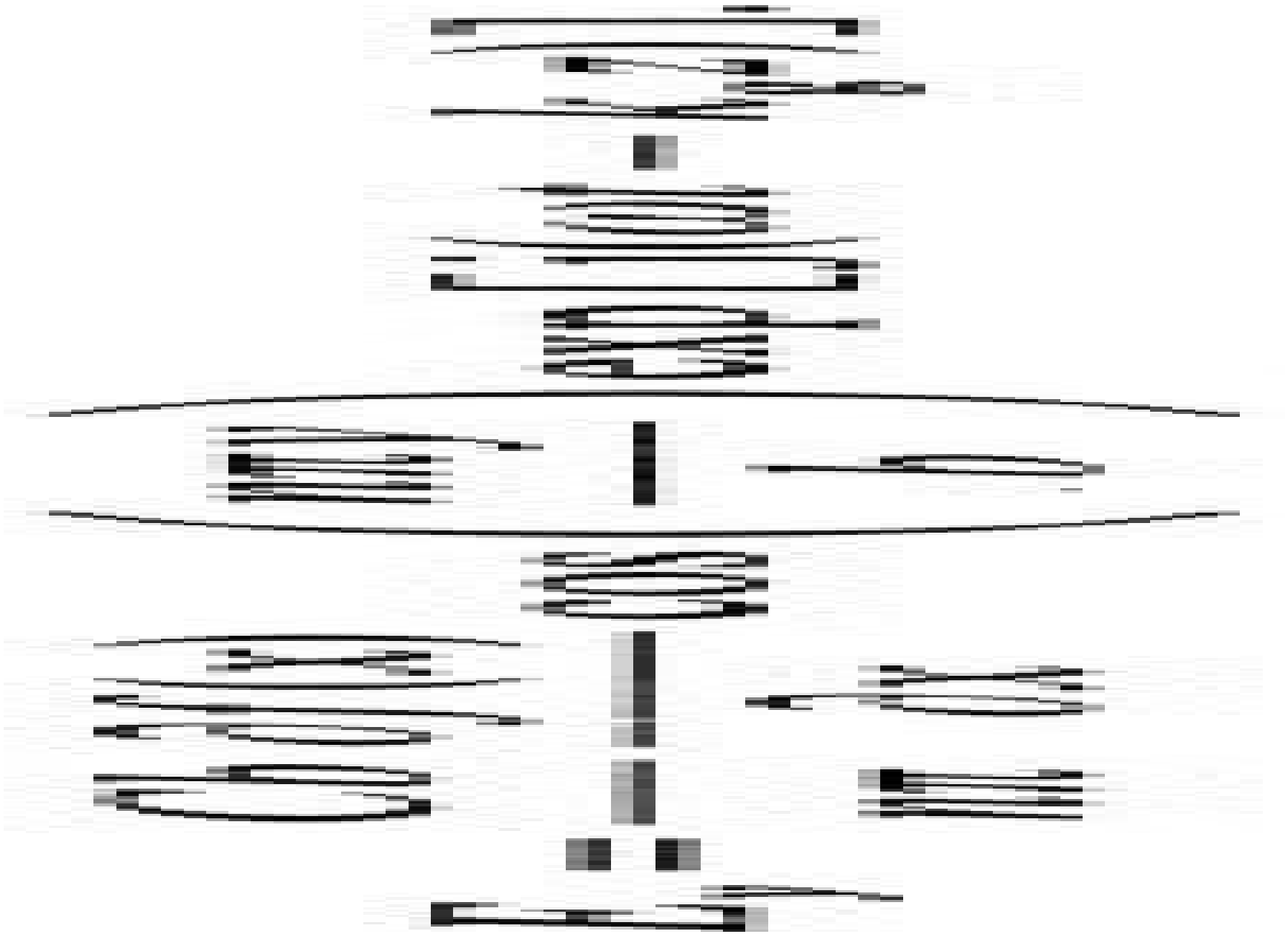


and the required y-component of the field is given by

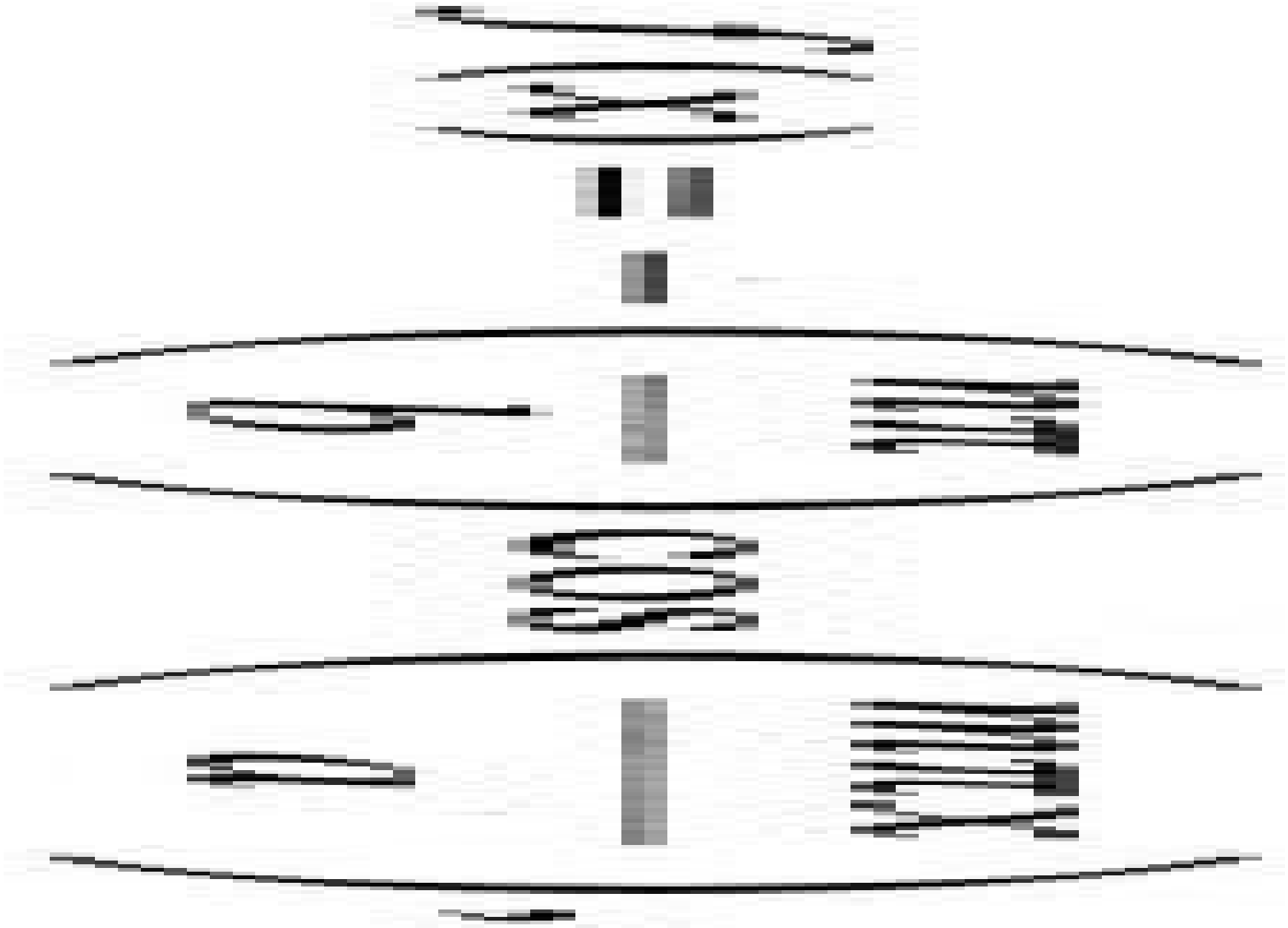


Integration with respect to y, putting the constant of integration zero, gives

(12.17)



We can now determine the function f(x) by making the component Ey satisfy the boundary condition that it be zero at the walls at x = 0 and x = a, and satisfy the wave equation. A function which does this, as can readily be verified, is



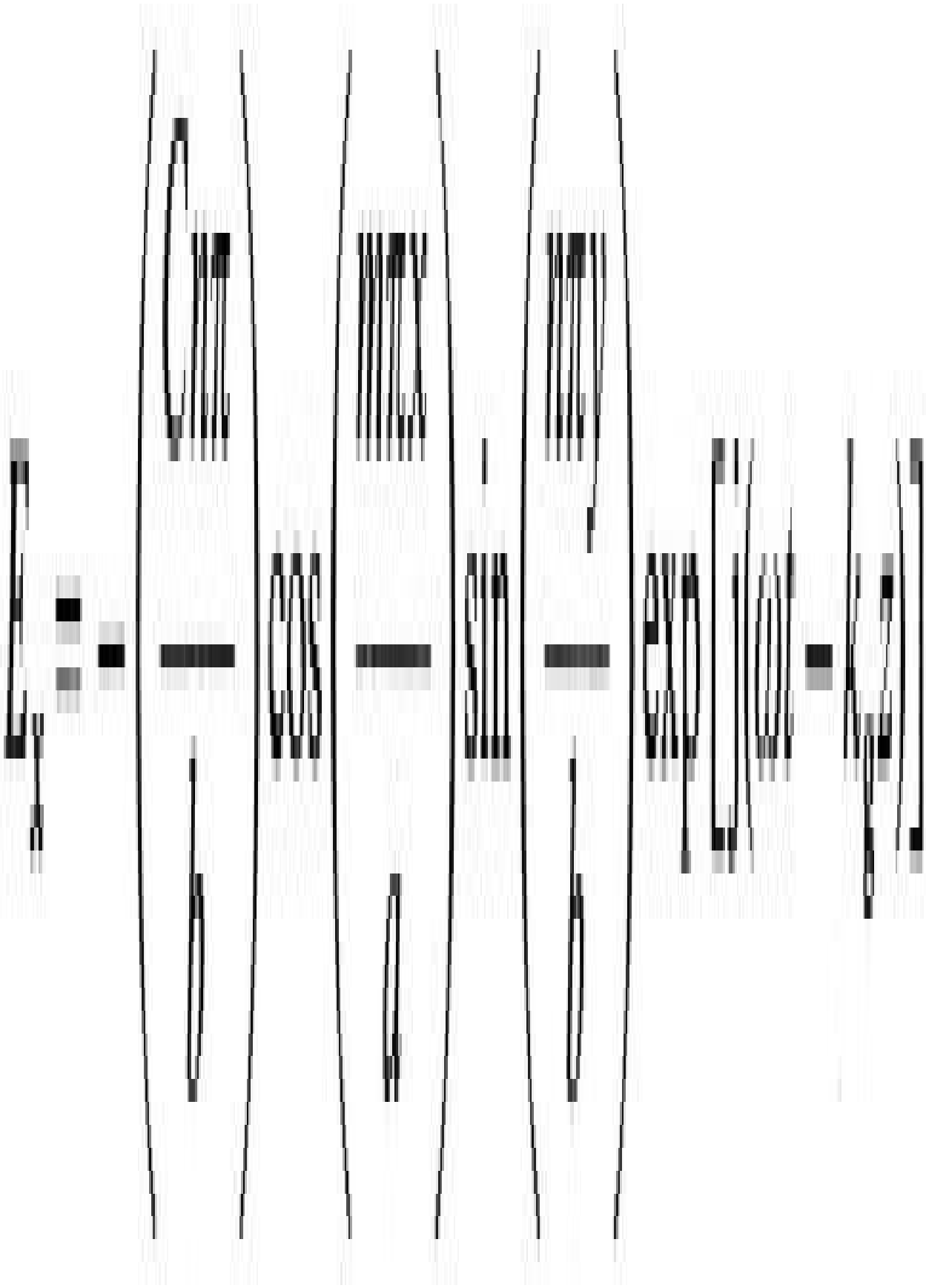
where m is zero or any integer. The constant ($n\pi/b$) has been chosen to make the expressions for the fields as symmetrical as possible. The parameters m, n and kg must obey the equation

(12.18)

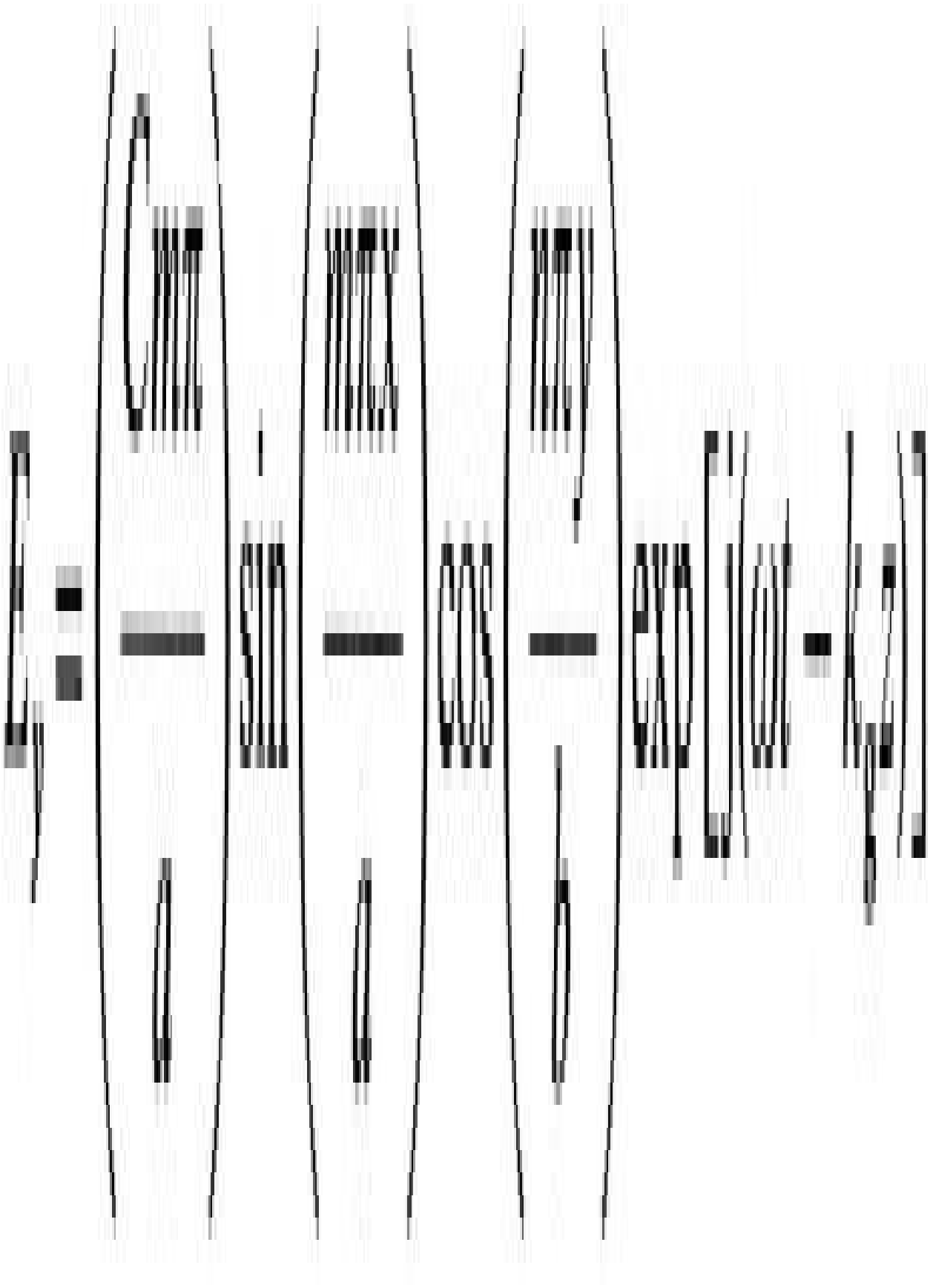
This last condition is obtained by substitution of Ey in the wave equation. It is called the waveguide equation, and gives the guide wavenumber kg of the allowed waves in terms of the integers m and n, the dimensions a and b of the waveguide, and the wavenumber k of the wave in free space. The guide wavenumber is related to the guide wavelength λg by the relation $kg = 2\pi/\lambda g$.

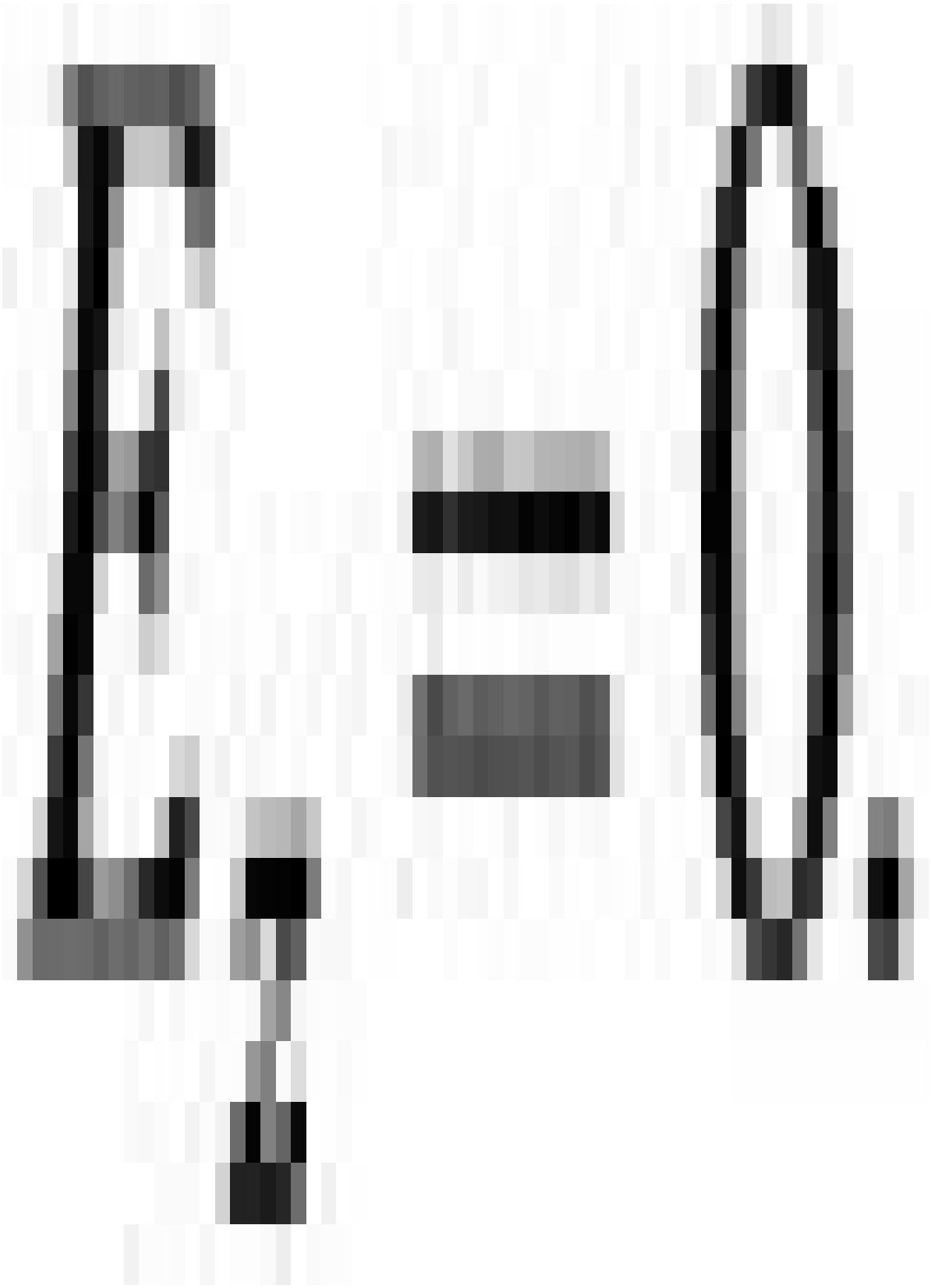
The expressions finally obtained for the electric field components in a TEmn wave are

(12.19)



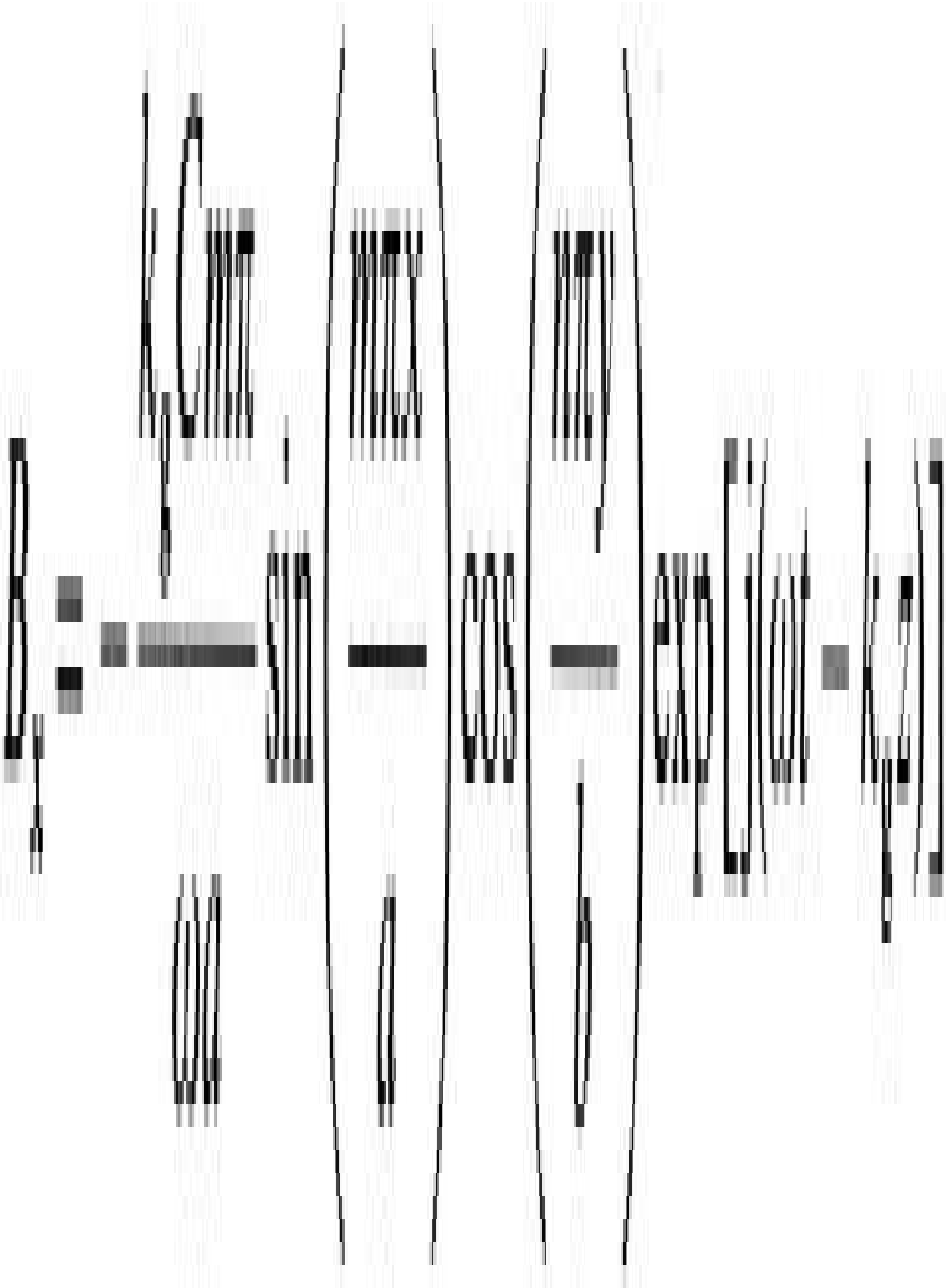
(12.20)



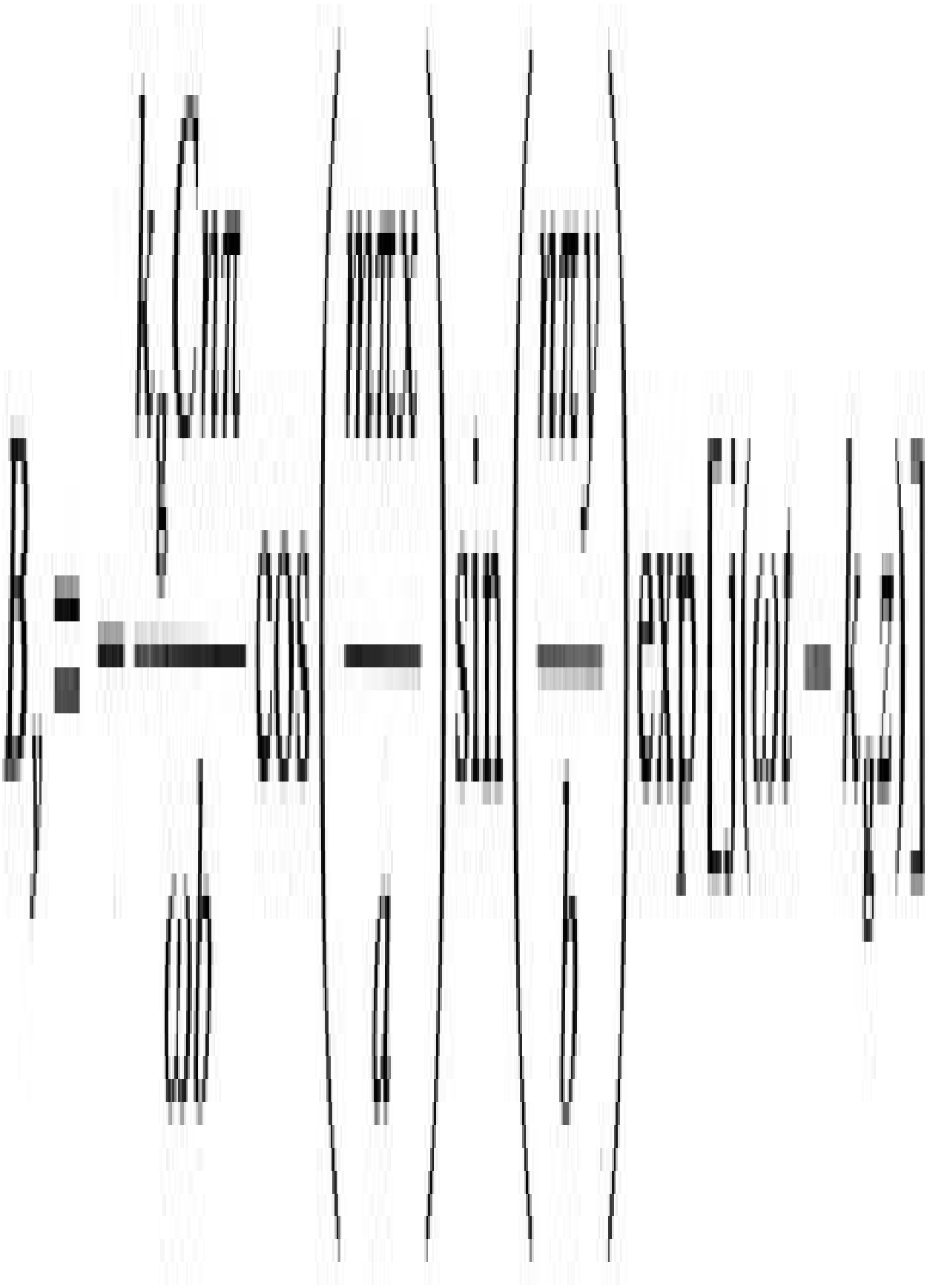


The magnetic fields can be determined in the usual way by using one of Maxwell’s equations relating the fields E and B. We obtain

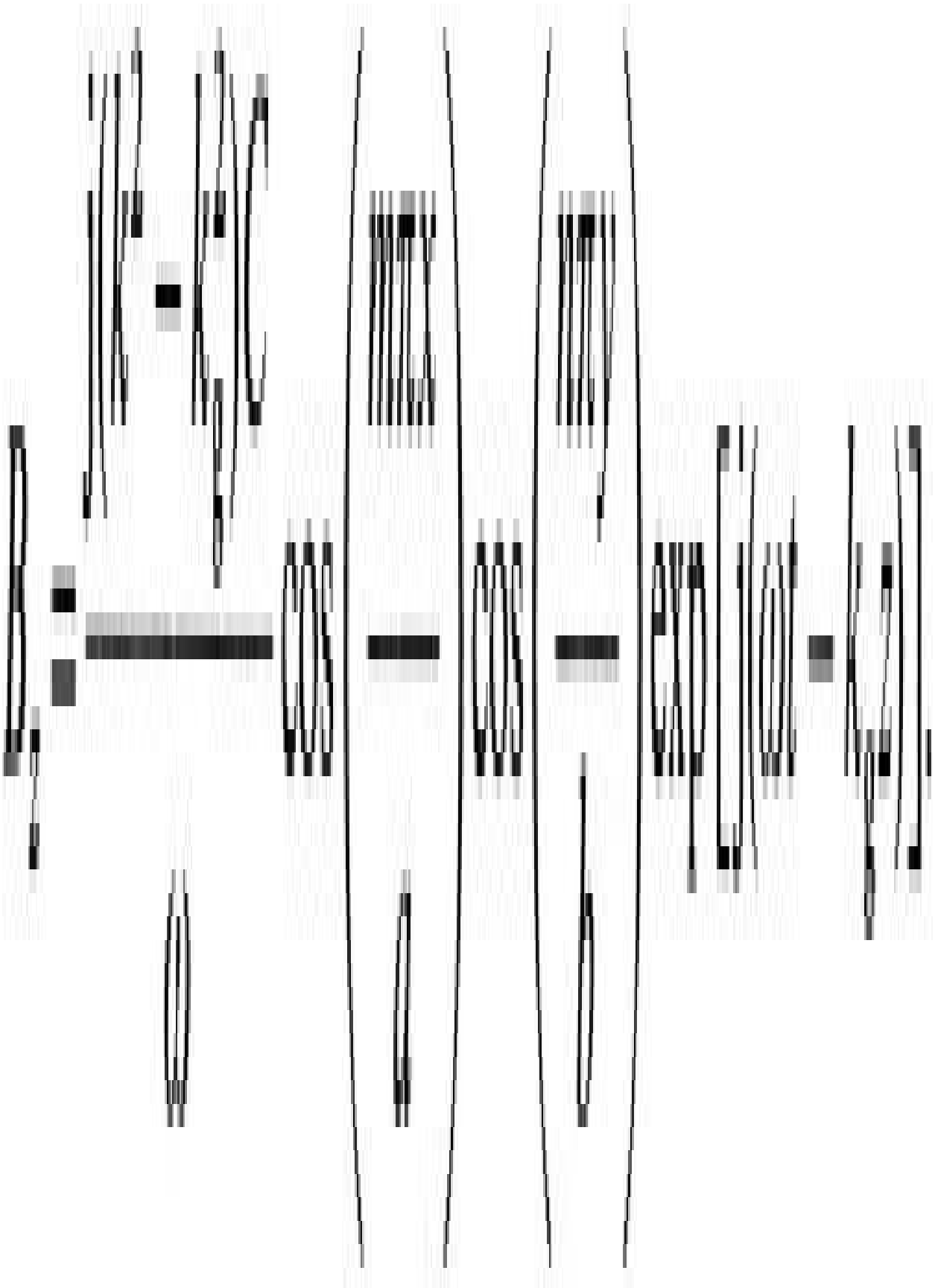
(12.21)



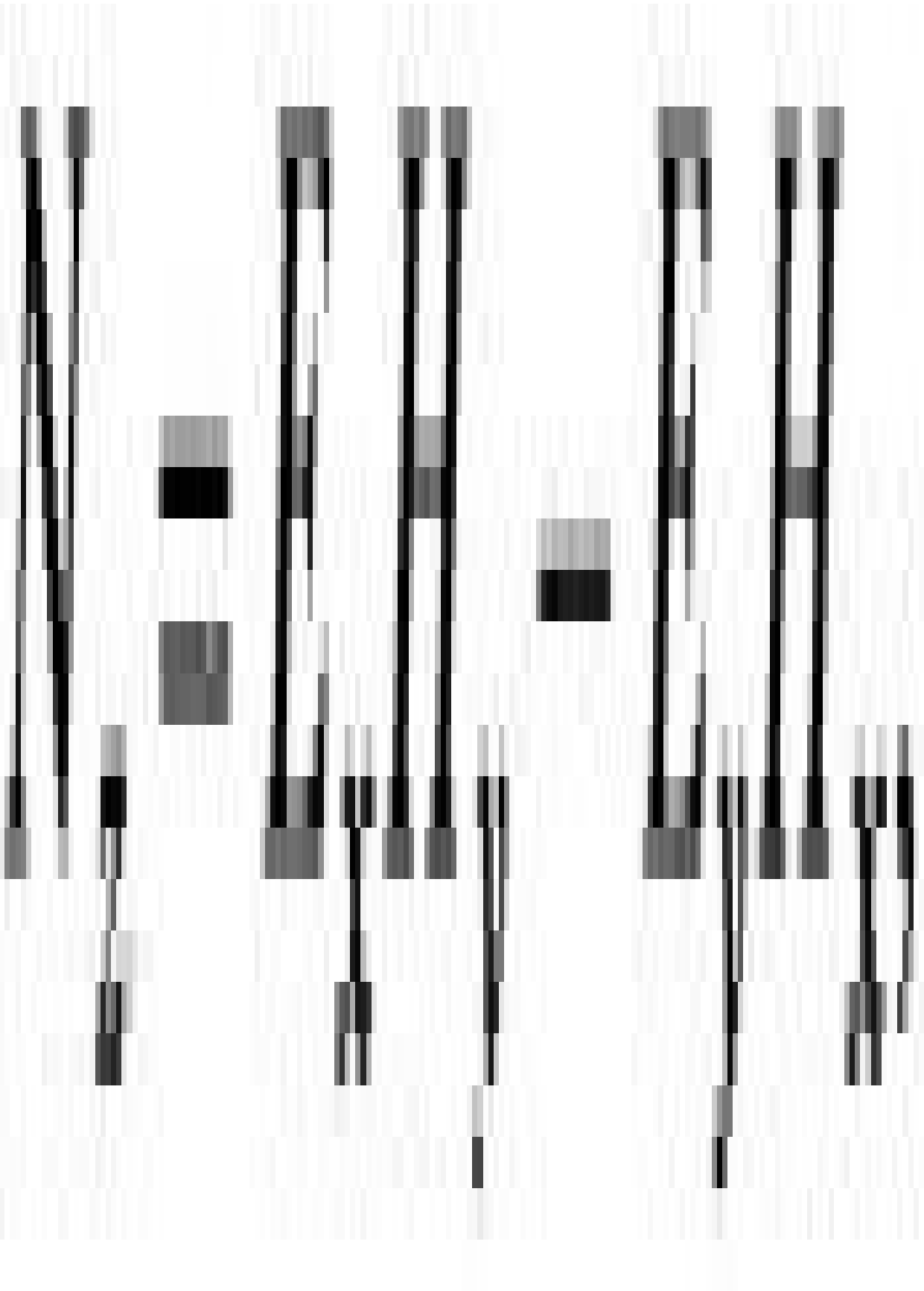
(12.22)



(12.23)

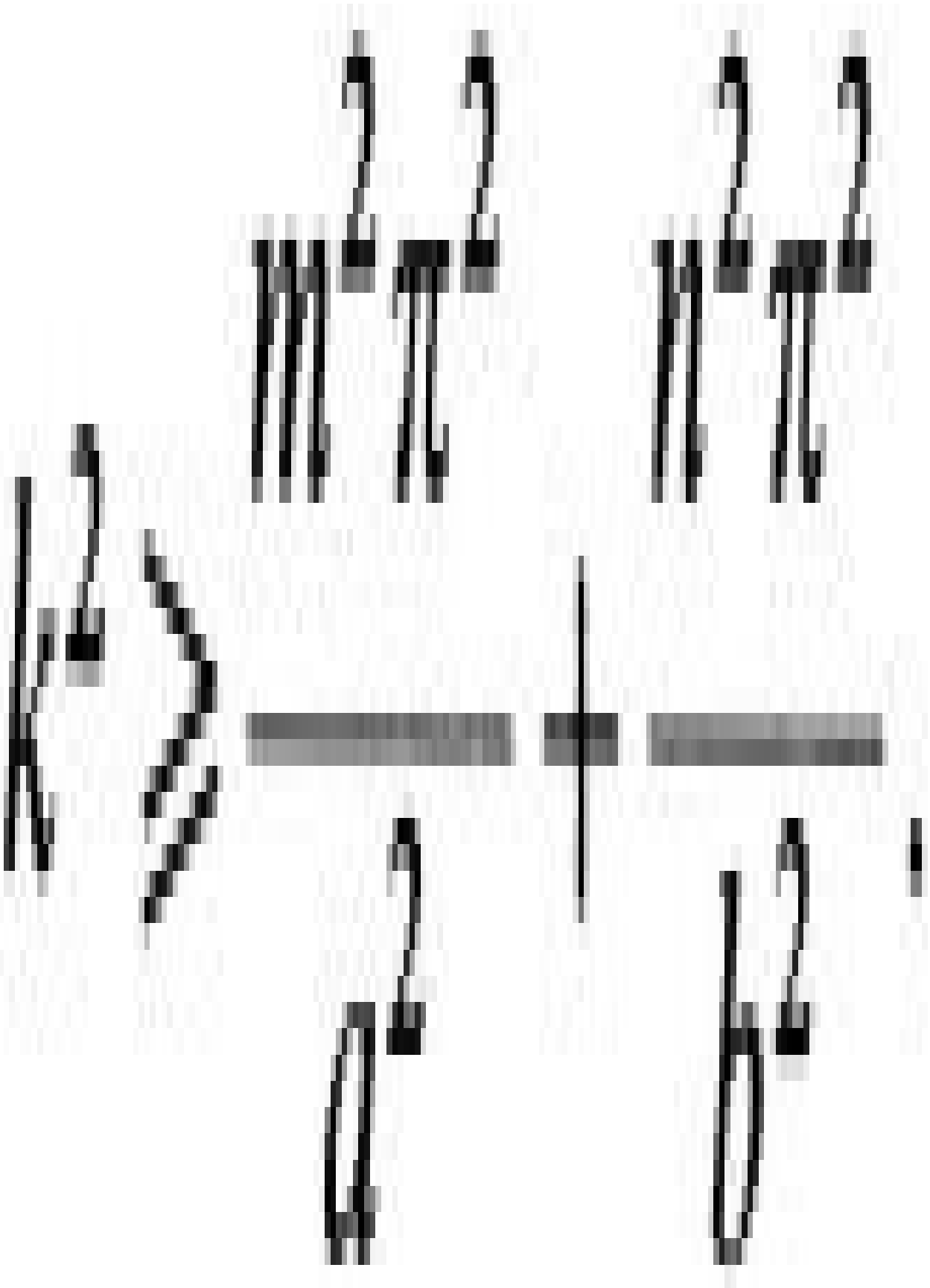


The above fields correspond to waves propagated in the z-direction, which is therefore the direction of the Poynting vector. Now



and together with the component Ex there goes an in-phase component Hy, and likewise for Ey and Hx. The component Hz is out of phase with Ex and Ey, and the energy flow in the x- and y-directions averages to zero.

From expressions (12.21) to (12.23) for the magnetic field, and expressions (12.19) and (12.20) for the electric field, it can be seen that for the wave to progress without attenuation in the z-direction, the guide wavenumber must be real. Otherwise the solutions obtained represent fields decaying exponentially with distance z. The waveguide Equation (12.18) tells us that kg is real only if



The different TEmn modes thus have different cut-off frequencies vmn below which a TEmn wave cannot be propagated in a guide of given size. The cut-off frequencies of the TE0n modes are the same as for waves between parallel plates. For the TEmn mode the cut-off frequency is given by

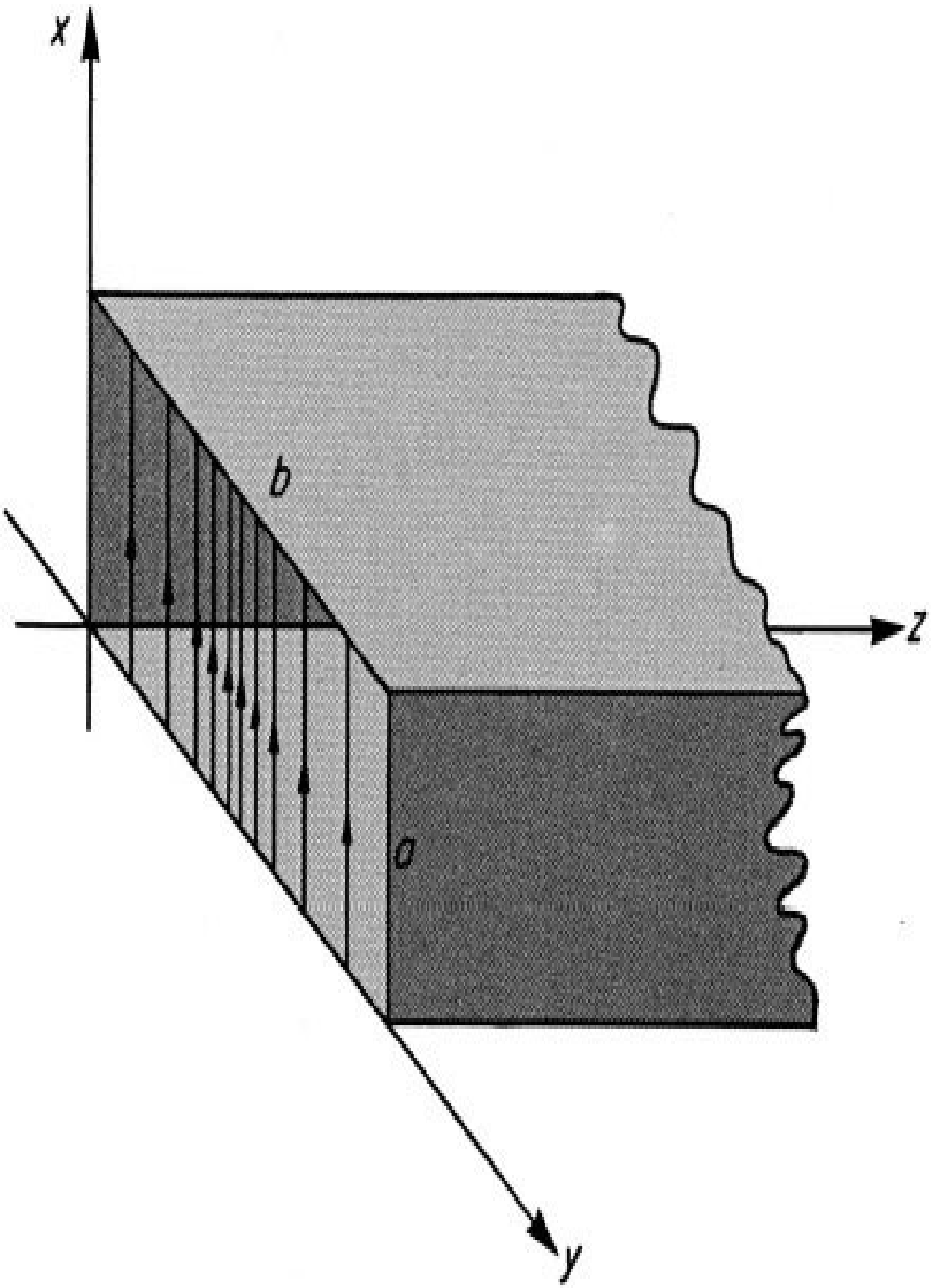
(12.24)

So far we have talked only of TE waves. There are also TM waves in which the magnetic field has no component in the z-direction, but in which the electric field has a non-zero z-component. The TM waves are also designated by integers m and n which characterize their behaviour in the x- and y-directions. They obey the identical waveguide equation to Equation (12.18) and have identical cut-off frequencies, given by Equation (12.24). There is one important difference however. There are no TM waves with either n or m equal to zero, i.e. no TM0n or TMm0 waves. Let us see why this must be so. Figure 12.5 shows the lines of the electric field in a TE01 mode in a waveguide at time t = 0 over the plane z = 0. The lines go from one wall straight across to the other, their density being greatest in the middle of the guide, as given by Equation (12.19). On the walls at x = 0 and x = a there are induced appropriate surface charges on which the lines of the field E begin and end. If a TM01 mode existed in which the magnetic field were independent of the x-coordinate the field B would also have to go straight across from one wall to the other. However there are no magnetic charges and the field B has to vary with the x-coordinate in order to make the lines of the field continuous. The lowest TM mode is thus the TM11 mode.

The TE and TM modes are the only solutions to Maxwell’s equations inside the guide that obey the boundary conditions. For radiation of a given frequency the general mathematical form of a wave travelling without attenuation in the z-direction is thus a mixture of all the possible TE and TM modes. The relative contributions to the total wave from the different modes is determined by the coupling of the waveguide to the source of radiation. In general the amplitudes of modes other than the mode with the lowest cut-off frequency are very small.

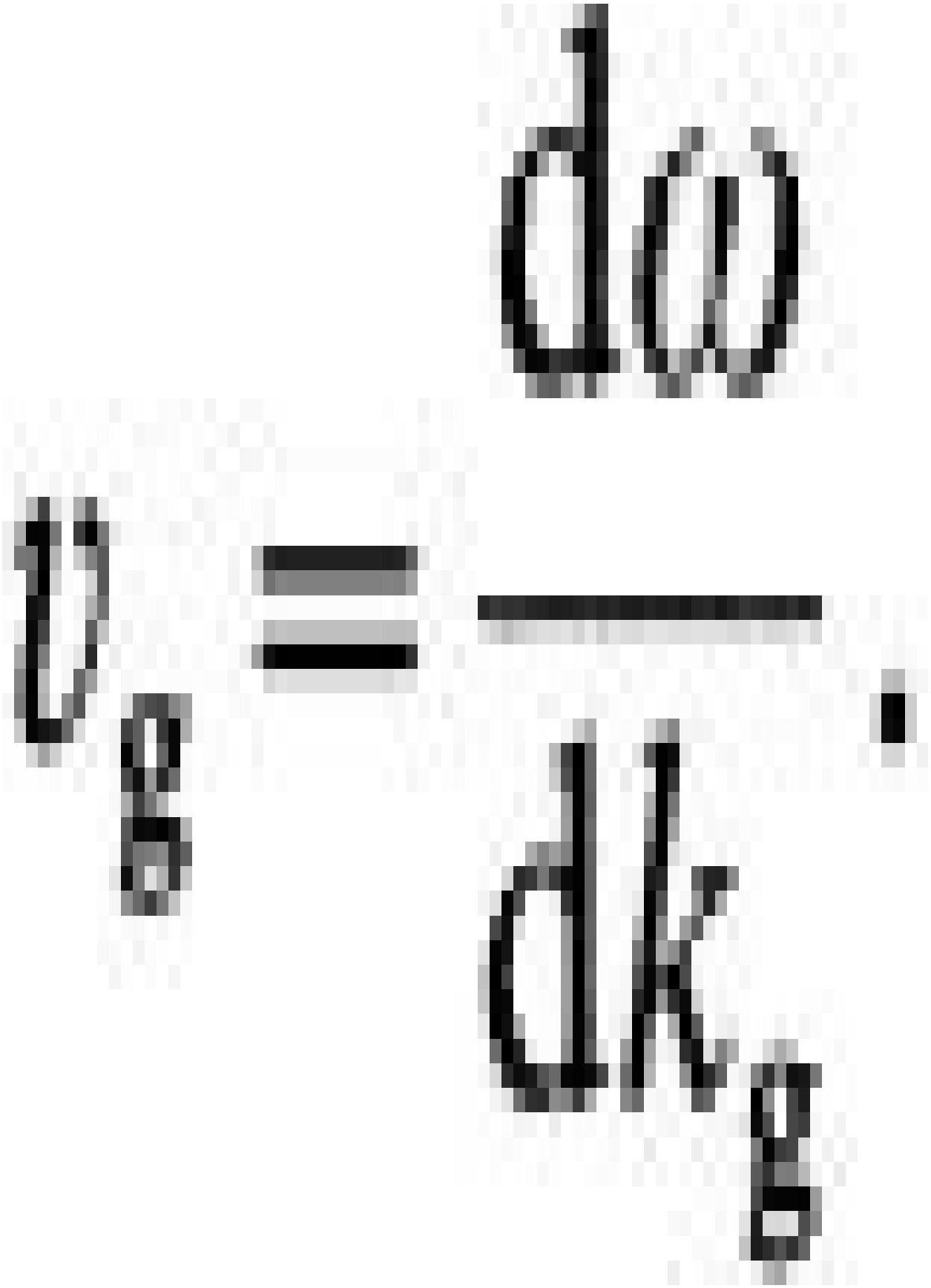
Higher modes are also more strongly attenuated in a real guide than is the lowest mode.

Figure 12.5. The lines of the electric field in the TE01 mode in a waveguide at time t = 0 over the plane z = 0.


x
b
z
a
y

12.2.1 The TE01 mode

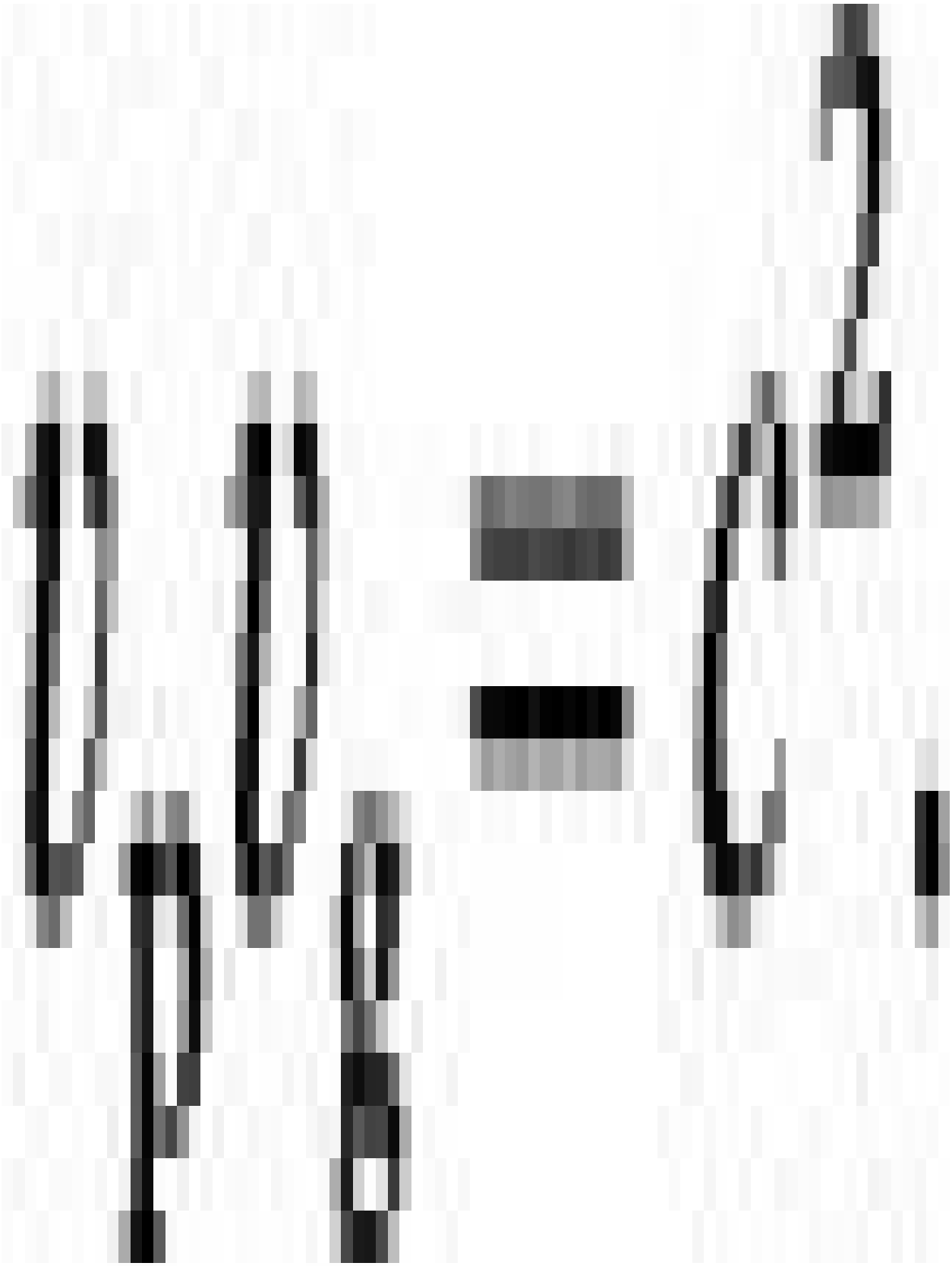
The lowest cut-off frequency of all is given by the smaller of the quantities c/2b or c/2a and corresponds to the electric field varying along the longest side of the guide. The TE01 mode shown in Figure 12.5 has the lowest cut-off frequency c/2b. The vanishing of the TM01 and TM10 modes has the important consequence that for a given frequency one can always choose a size of waveguide such that only the lowest TE wave is propagated. In practice waveguides are usually used under those conditions. Consider 3 cm wavelength radar waves as an example; their frequency is 10^{10} Hz. If a guide has dimensions 2 cm one way (b = 2 cm) and 1 cm the other, Equation (12.24) gives the following values for the cut-off frequencies of the different modes: $v01 = 7.5 \times 10^9$ Hz; $v10 = v02 = 1.5 \times 10^{10}$ Hz; $v11 = 1.68 \times 10^{10}$ Hz; etc. The lowest TM mode, the TM11, has a cut-off frequency $v11 = 1.68 \times 10^{10}$ Hz. Since only the TE01 mode has a cut-off frequency lower than the frequency of the radiation, only the TE01 mode can be propagated. The guide wavenumber kg of the TE01 mode in the waveguide is given by Equation (12.18) to be about 1.4 cm^{-1}. The wave velocity vp of the wave is ω/kg which has the value $\sim 4.5 \times 10^8$ m s^{-1}. The speed with which the energy in the wave travels down the guide is given by the group velocity of the TE01 wave obtained from Equation (11.28)



This is evaluated in Equation (12.14) as

(12.25)

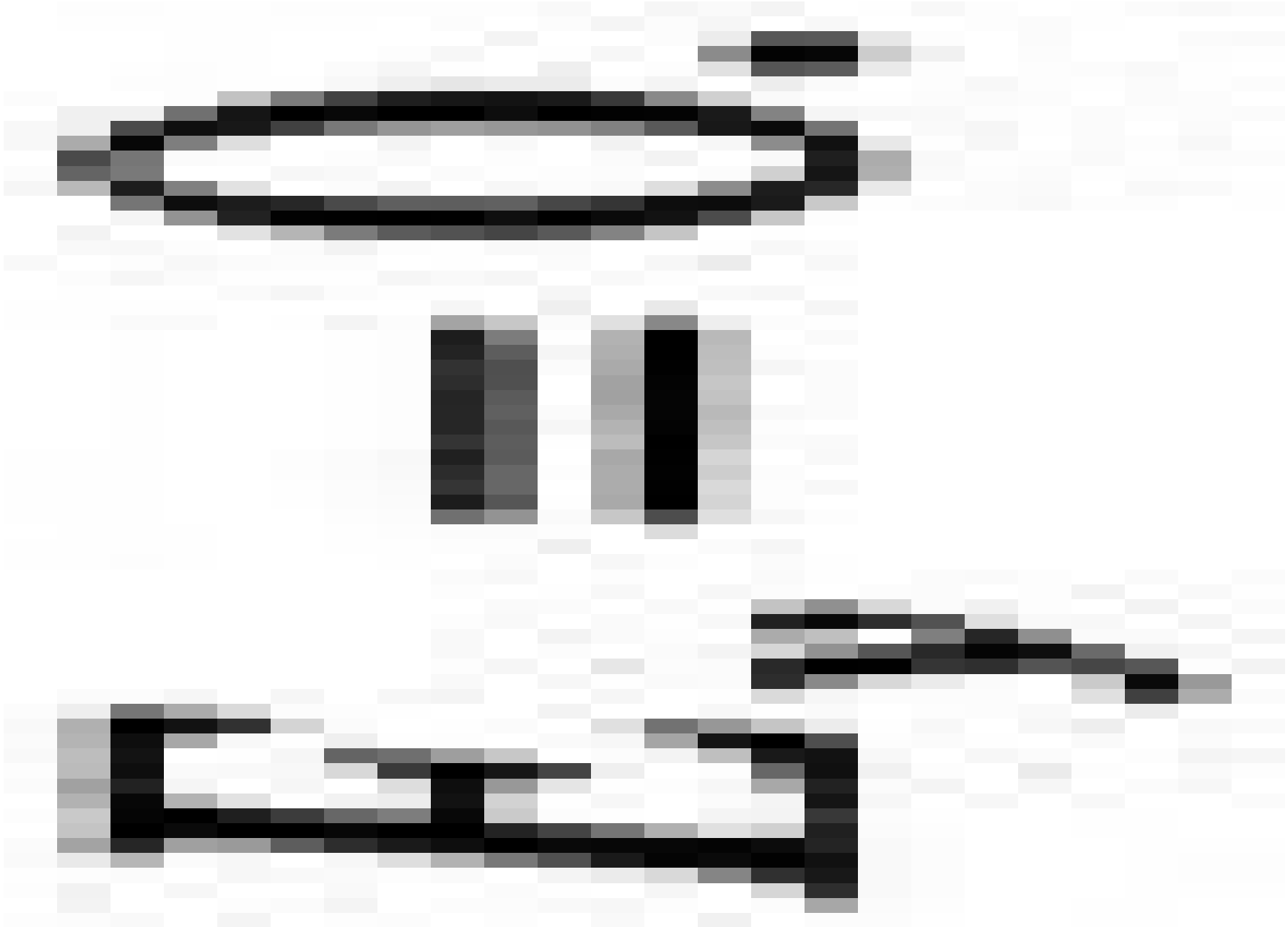
and the product of the wave and group velocities is equal to the square of the speed of light in free space (Equation (12.15))



The energy in the TE01 mode in our example thus flows down the guide at a speed of about 2×10^8 m s−1.

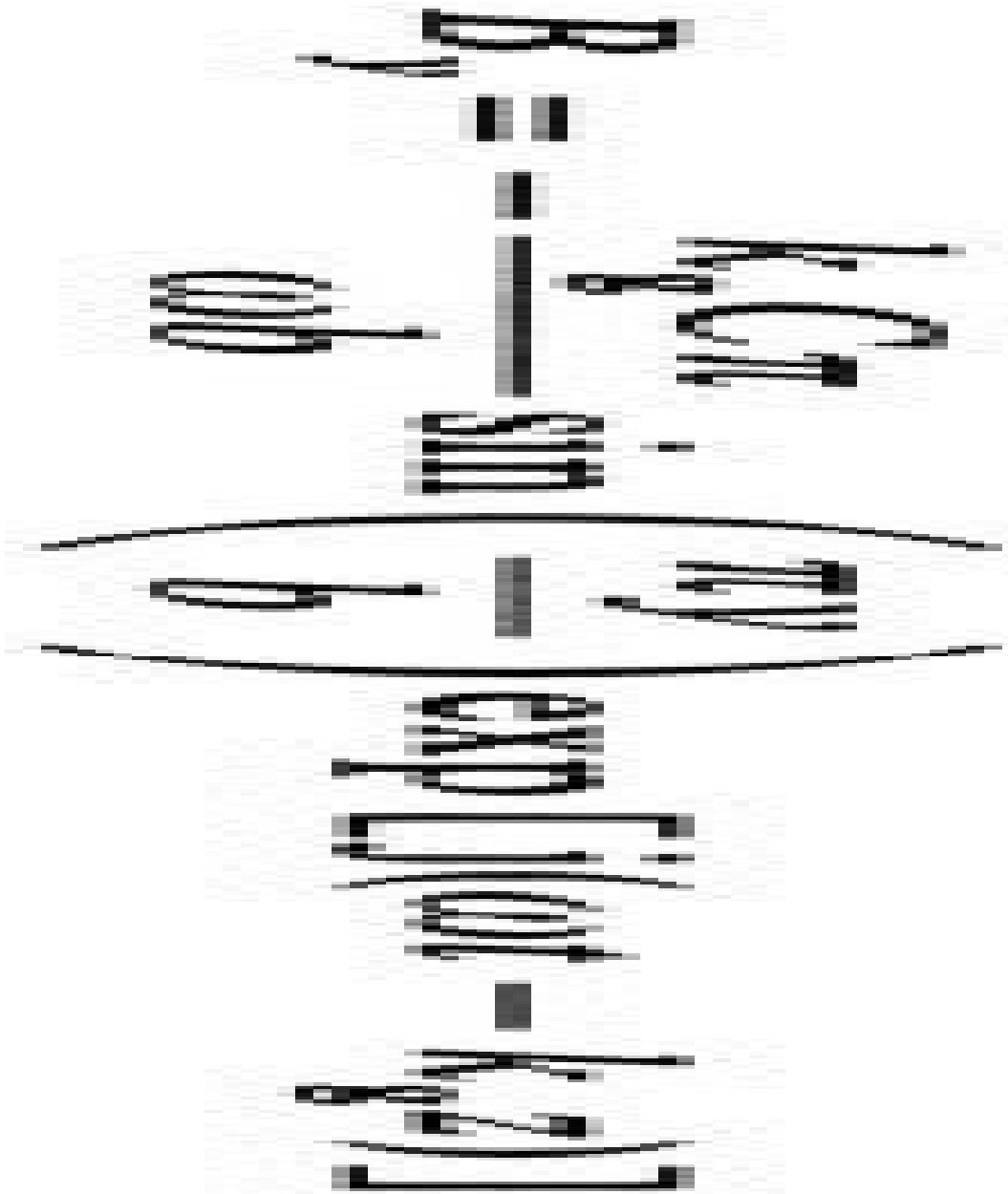
The TE01 wave is the most important and most commonly encountered wave in a rectangular guide. The electric field in this wave is given by substituting m = 0, n = 1 in Equations (12.19) and (12.20), and we obtain

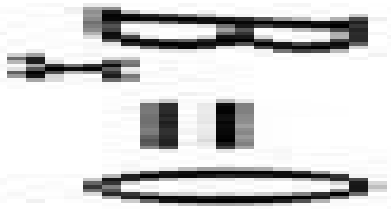
(12.26)



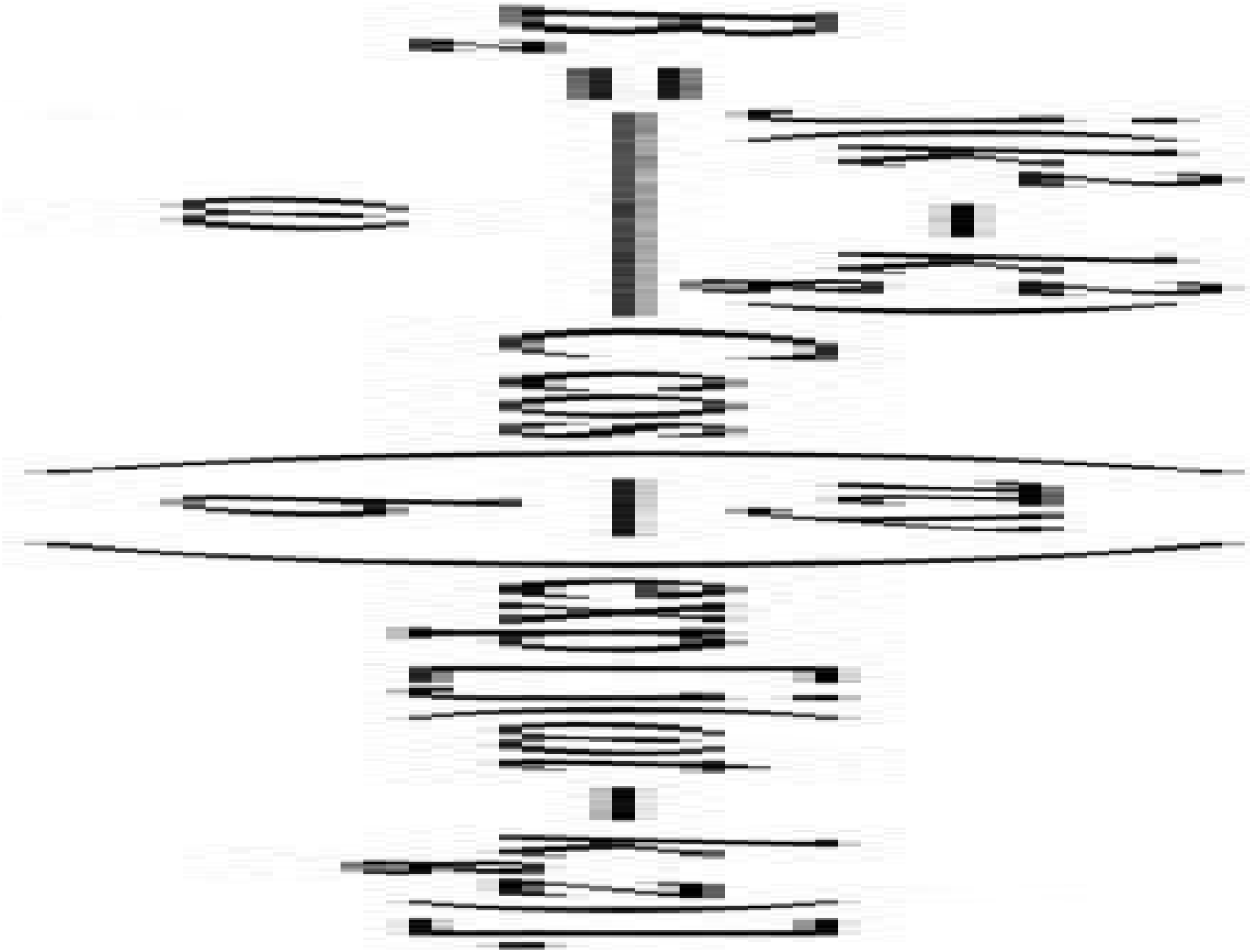
This field has already been shown on Figures 12.2 and 12.3. The magnetic field in the TE01 wave can be obtained by substituting m = 0, n = 1 in Equations (12.21) to (12.23), to give

(12.27)



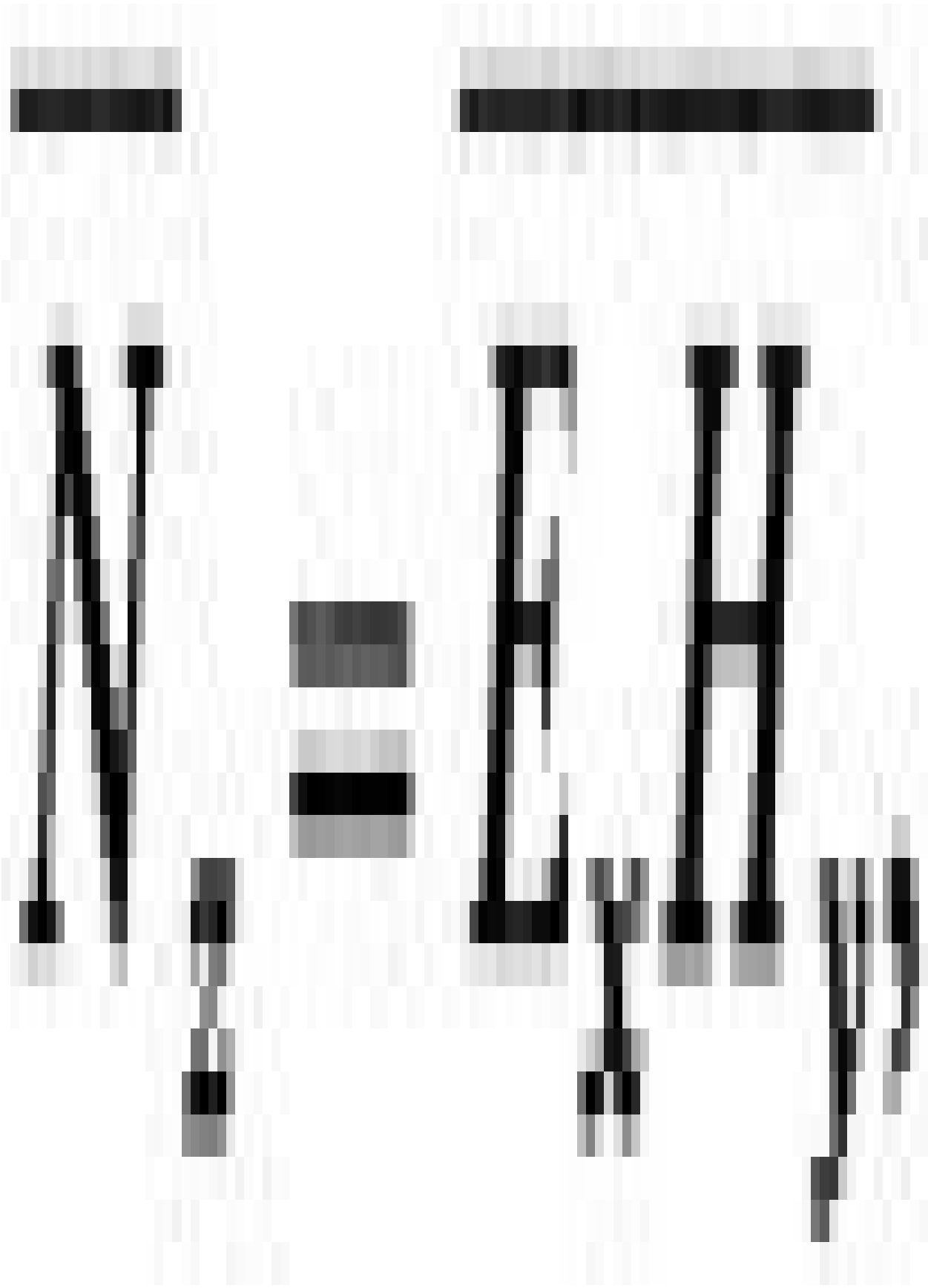


(12.28)

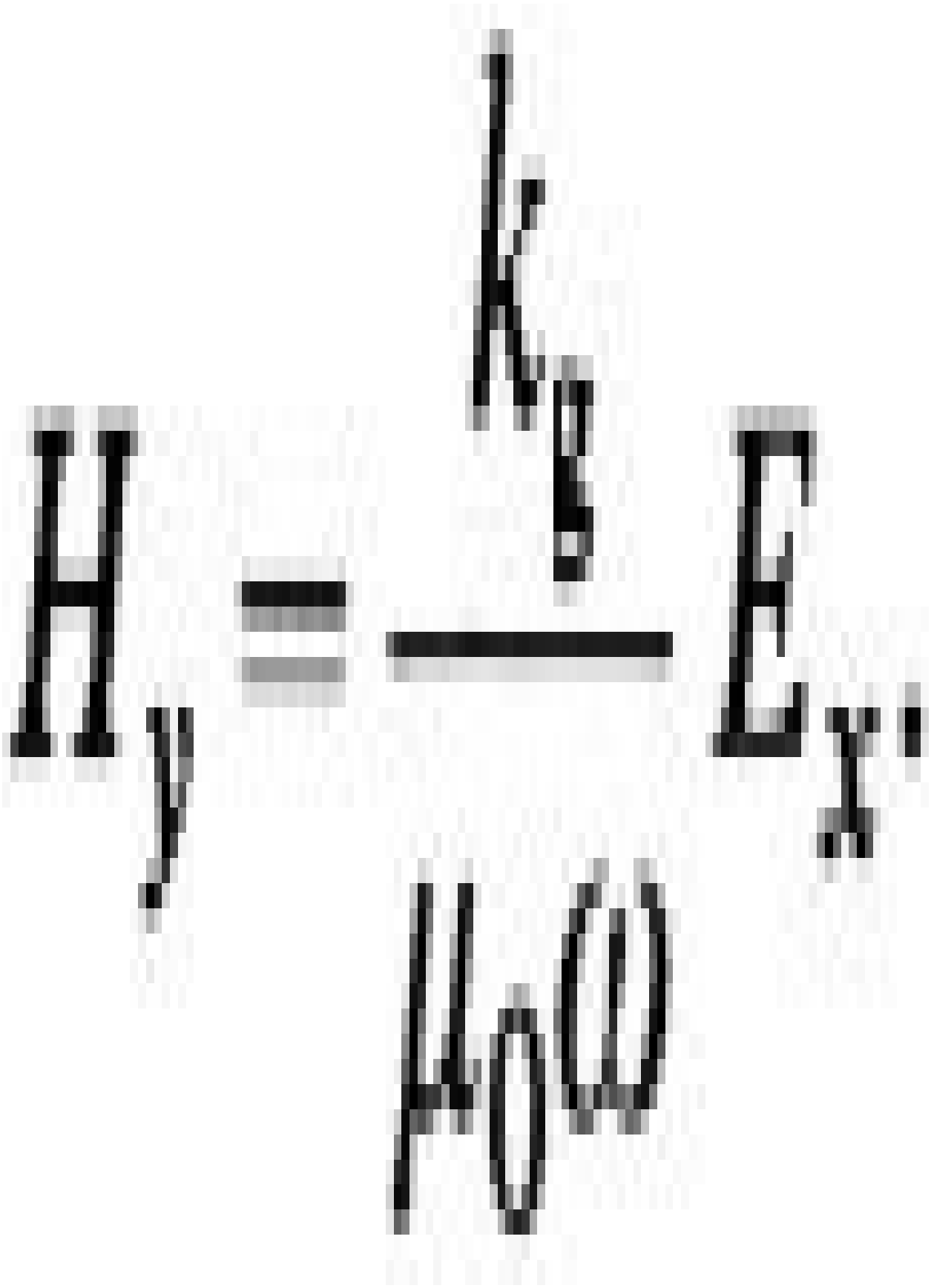


The magnetic field described by these equations at time t = 0 is shown in Figure 12.6. The field lines have been drawn for the plane x = a/2 only. Since the field is independent of the coordinate x, the field lines look the same in any plane parallel to the one drawn.

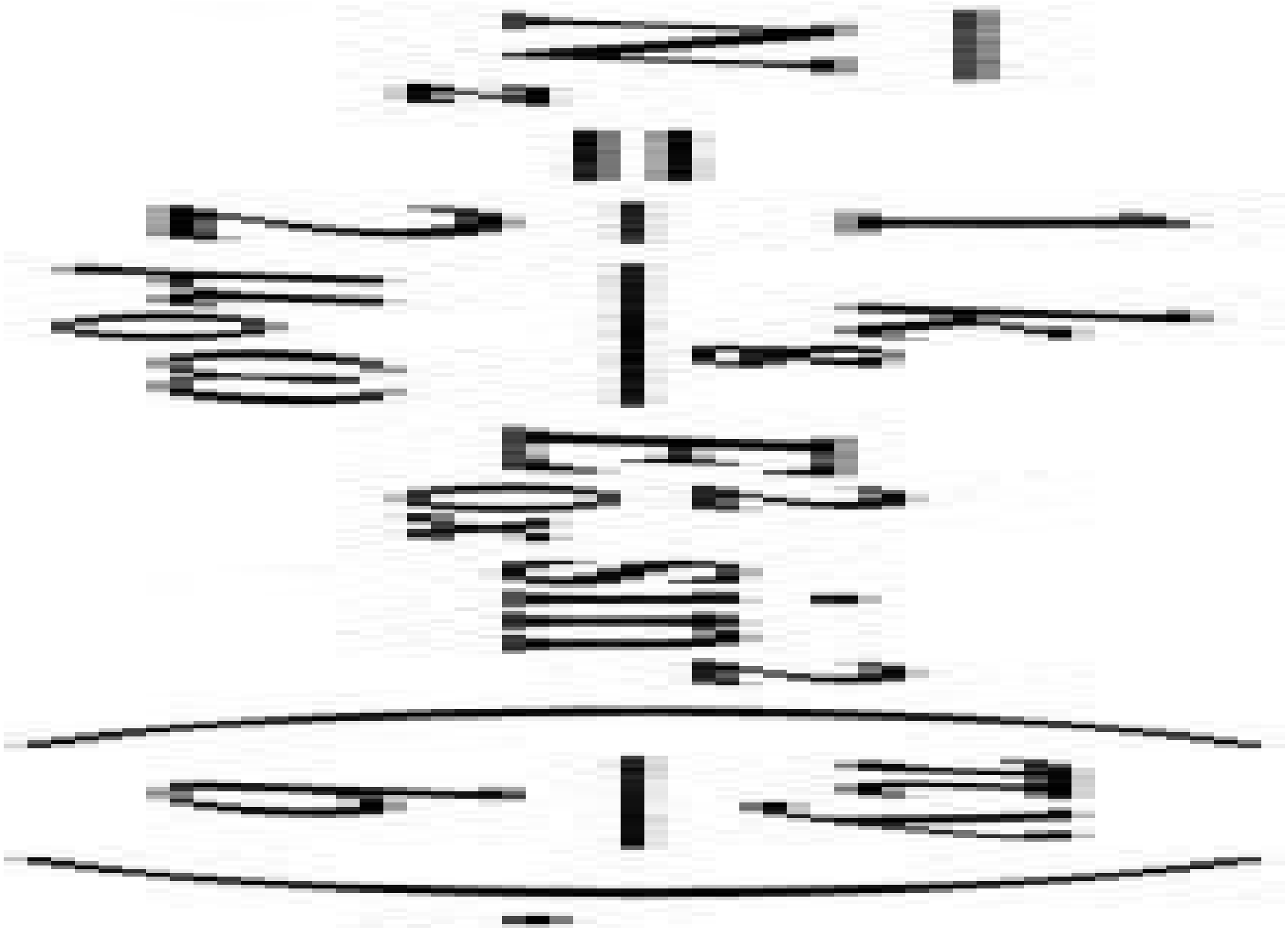
The electric field in a waveguide can be very large. A ship's radar set typically works at about 0.5 MW power. (Of course the average power delivered by the source of waves is very much less than this, since the source is on for only a small fraction of every second.) Thus the time average of the Poynting vector integrated over the cross-sectional area of the guide must equal 0.5 MW during operation. The Poynting vector for the fields in the waveguide is in the z-direction, and for the TE01 mode



where the bars over the symbols indicate averages over a period of the oscillation. From Equations (12.26) and (12.27) we see that the y-component of the magnetic intensity is

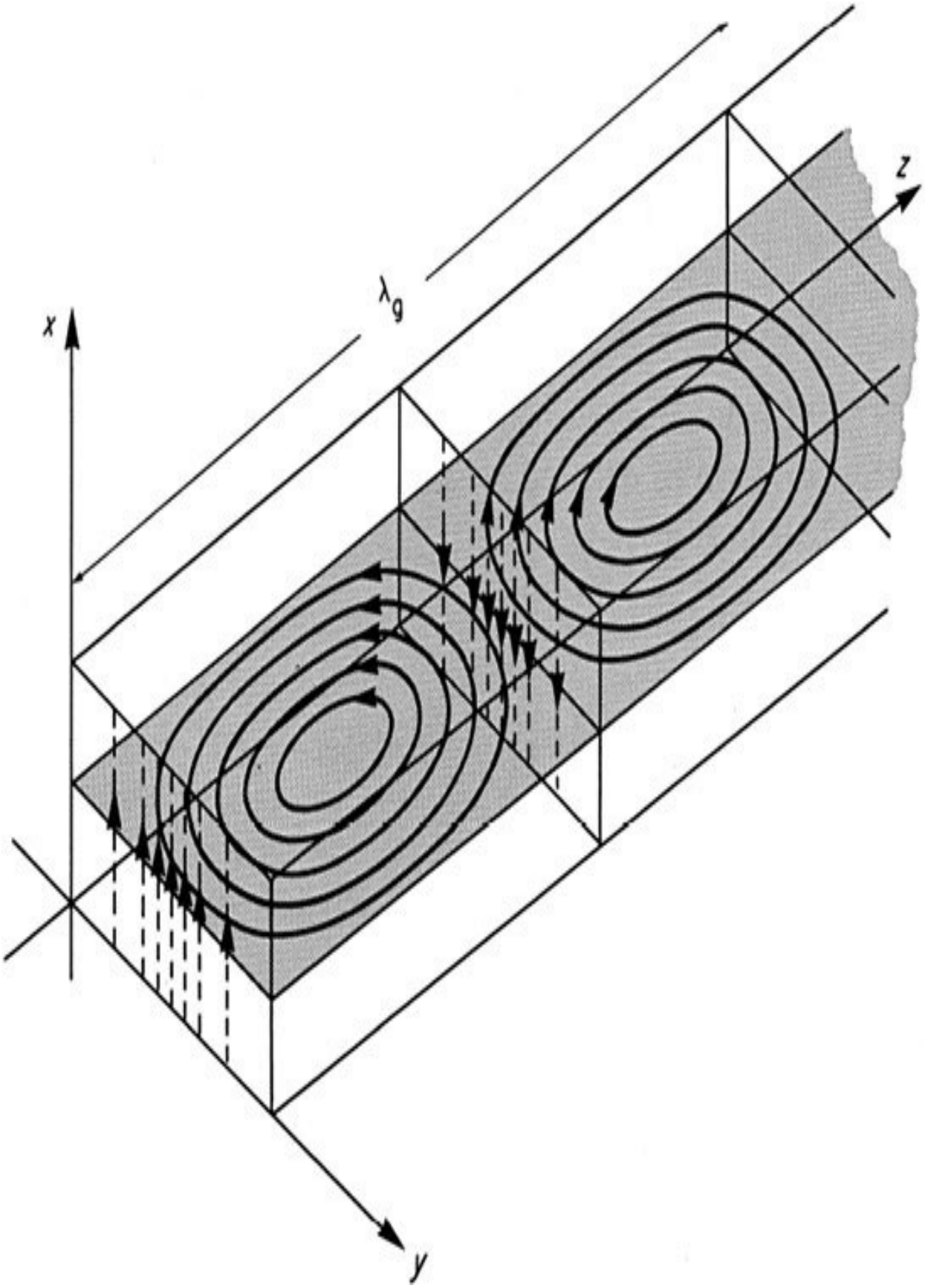


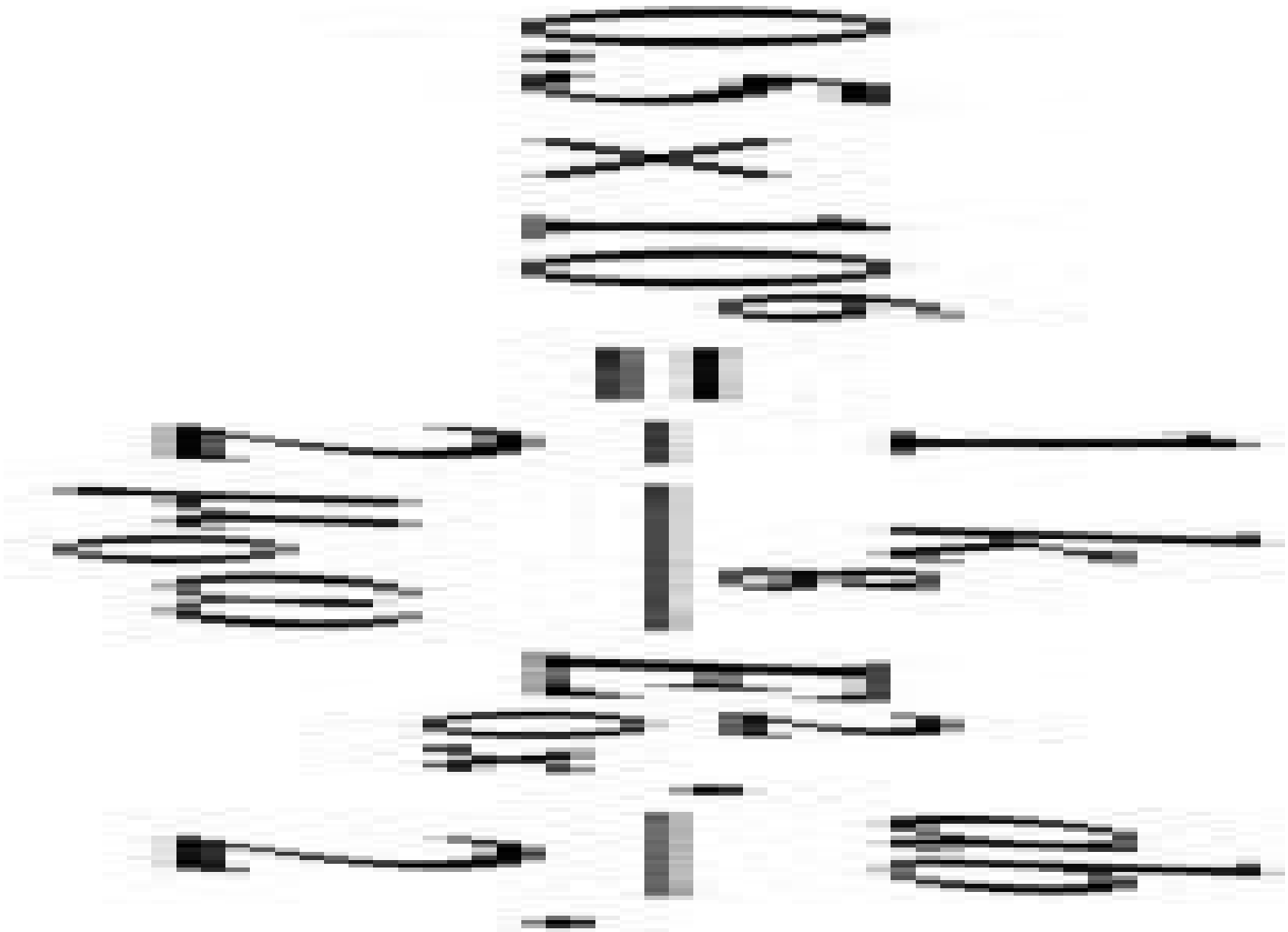
Hence if E0x is the maximum amplitude of the component Ex in the guide



Integrating this expression over the cross-sectional area of the guide and equating the result to the total energy flow per second, we have

Figure 12.6. The magnetic field lines in the TE01 mode in a rectangular waveguide. The field lines are drawn at a time when the electric field, shown dotted, is a maximum in the + x-direction over the plane z = 0. The magnetic field lines look the same in any plane parallel to the plane x = a/2 drawn.


z
λ_g
x
y



On substituting into this expression the typical values a = 0.01m, b = 0.02 m, $\omega = 2\pi \times 10^{10}$ s−1 and kg = 1.4×10^{2} m−1 we obtain the value of ~ 2.4 MV m−1 for the maximum amplitude E0x.

Waveguides commonly in use do not have superconducting walls, and so there are losses, as mentioned in the introduction to this chapter. The energy in a wave decreases as the wave propagates down the guide, due to heating of the guide walls. The heating arises because of the currents induced in the walls by the wave, which penetrates to a depth of the order of the skin depth δ (discussed in section 11.5). The average heating loss in a small length δz of the guide can be estimated by assuming that the currents in the walls flow only in a thin layer of thickness δ. The currents can be determined by using Ampère's integral law for the field H (Equation (5.29)), and hence the heating losses estimated in terms of the conductivity σ and skin depth δ of the walls. Comparison of the heating loss in the length δz with the energy flowing down the guide at the coordinate z, then gives an estimate of how rapidly the energy is attenuated (see Problem 12.6). For a typical brass rectangular waveguide, with sides 2 cm and 1 cm wide, the energy of a TE01 wave of frequency 10^{10} Hz falls to one half over a distance of about 16 m.

12.2.2 Further comments on waveguides

There is usually a reflected wave, however small in amplitude, whenever there is a discontinuity in a waveguide or by whatever means it is terminated. A guide terminated by a load consisting of absorptive material of gradually increasing volume over several wavelengths has what is called a matched load and such a termination results in almost no reflected wave. If the guide is coupled to an antenna designed to radiate the wave into free space, the wave reflected off the specially designed endpiece is very small. If the waveguide is terminated by a metal sheet the reflected wave has an amplitude very nearly equal to the amplitude of the incident wave. The reflected wave combines with part of the forward going wave to produce a standing wave pattern, analogous to standing

waves on a transmission line. The standing wave pattern in a waveguide can be used to measure the amplitude of the reflected wave, and may be investigated by inserting a probe into the interior of the guide through a slot in a wall. (See Problem 12.9.) The slot should be placed where it will not interfere with the surface currents in the wall, so that the wave in the guide suffers minimum interference. The surface currents in the walls are perpendicular to the magnetic fields lines. For the TE01 mode the magnetic field lines are drawn in Figure 12.6 and the magnetic field gives rise to the surface current distribution over the top broad wall shown on Figure 12.7. A narrow slot along the z-axis in the middle of the wall will thus have minimum influence on the wave propagation. Figure 12.8 shows an arrangement often used to measure the standing wave ratio. A crystal diode rectifies the voltage produced between a point on the probe and earth. The diode gives a steady voltage out which is roughly proportional to the square of the amplitude of the electric field in the middle of the guide.

Figure 12.7. At times when the fields in the waveguide are as shown on Figure 12.6, the surface currents over the top face are as shown above. In the centre of the top face the currents are parallel to the z-axis, and a slot cut down the guide in the middle will have little effect on the distribution of surface currents and hence fields in the guide.

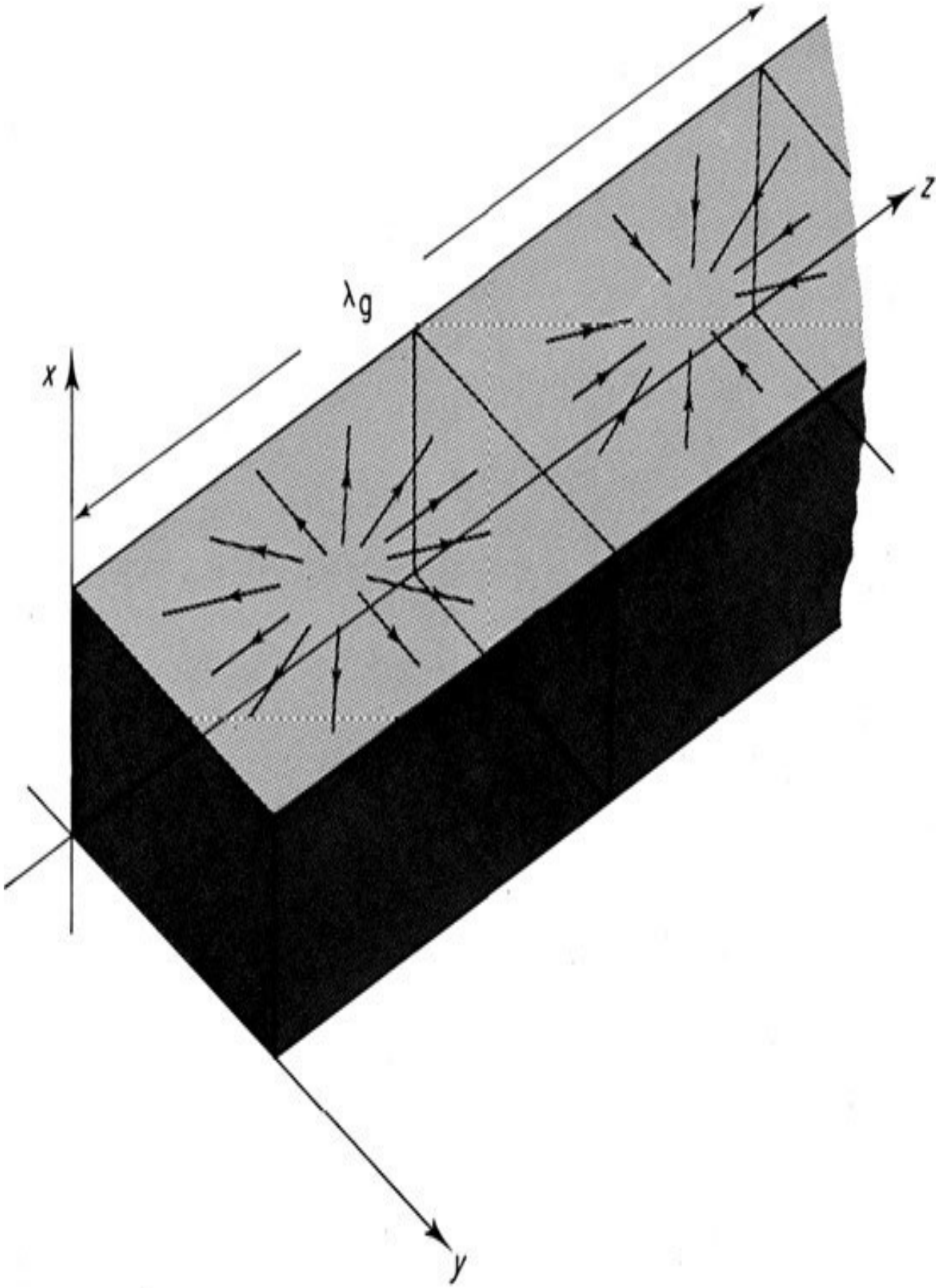
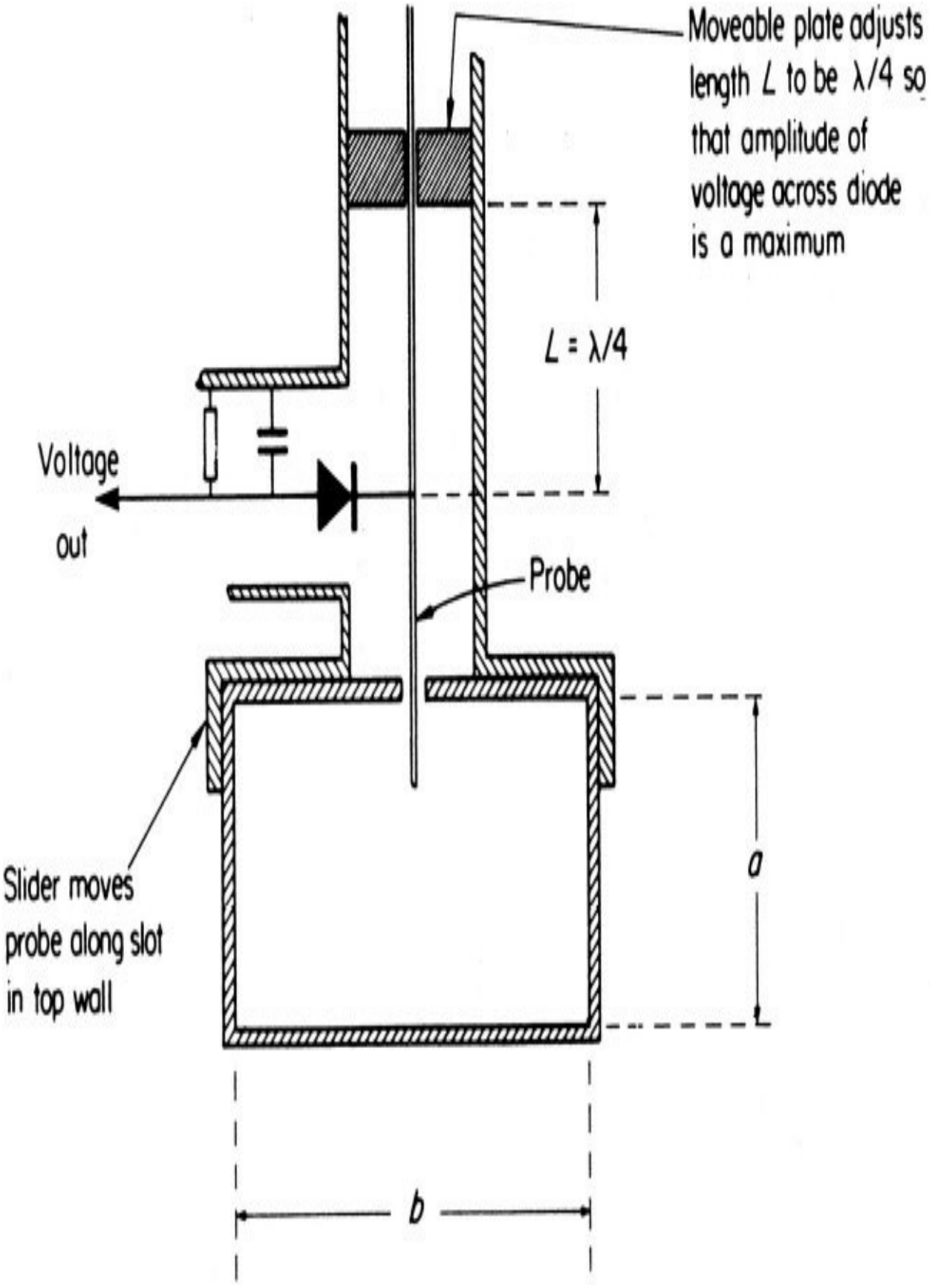

λ_g
x
z
y

Figure 12.8. A crystal detector in a slotted guide.


Moveable plate adjusts length L to be λ/4 so that amplitude of voltage across diode is a maximum
L = λ/4
Voltage out
Probe
Slider moves probe along slot in top wall
a
b

The group velocity of any TEmn mode can be found in the same way as that for the TE01 treated in the last section. It is a straightforward matter to show that Equation (12.25) is true for all modes. Thus it is generally true that the product of wave and group velocities for any mode is equal to the square of the speed of light in vacuo, although the separate wave and group velocities vary from mode to mode.

The dispersion in waveguides also depends upon mode. The amount by which the wave velocity vp changes with frequency can be determined from Equations (12.11) and (12.18), and depends upon the mode integers m and n. Since guides are rarely used in any but the TE01 mode this is usually of little practical importance. Dispersion in the TE01 mode can influence the time structure of pulses of radiation sent down a guide. The pulse corresponds to a group of waves in a band of frequencies, and the low frequencies in the group travel down the guide slower than the high frequencies. This results, as is usual in dispersive situations, in the pulse spreading out in time as it proceeds down the guide.

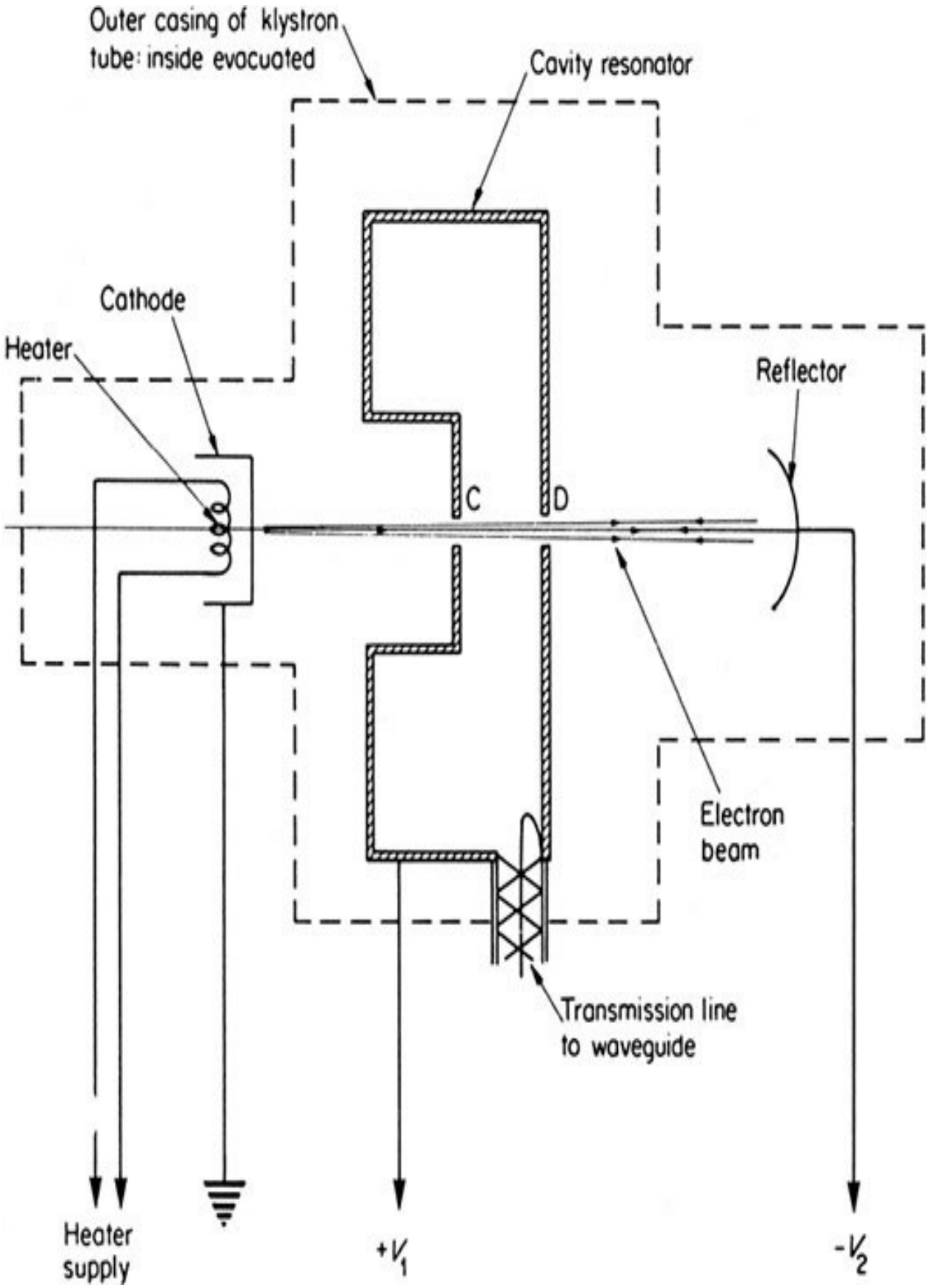
Waveguides with circular cross-sections are used in special applications, such as transmission of circularly polarized waves. The solutions of the wave equation which obey the boundary conditions at the walls again correspond to TE and TM modes characterized by integers m and n. The mathematical description of the modes is more complex than for rectangular waveguides and we will not go into it here. Optical fibres provide an interesting example of the transmission of electromagnetic waves down circular waveguides. An optical fibre is a thin thread of transparent dielectric, with diameter usually of a few up to a few hundred microns, enclosed in a cladding of material of lower refractive index. The fibre transmits radiation of optical frequencies over large lengths with little attenuation. If the diameter of the fibre is much greater than the free space wavelength of the light to be transmitted the propagation may be treated by the ray tracing techniques of geometrical optics. A ray travelling down the fibre at a small angle to the fibre axis is totally internally reflected each time it strikes the fibre surface, and can suffer several thousand reflections per metre with sufficiently small loss that optical signals can be sent along fibres over distances

of hundreds of kilometres and retain sufficient strength to be received and interpreted. If the diameter of the fibre is comparable to that of the light to be transmitted, further insight into the propagation may be obtained by treating it in terms of modes propagating in a circular waveguide. If the fibre diameter and light wavelength have appropriate values, only the lowest mode will propagate, as in rectangular guides. At optical frequencies this would happen if the diameter were about a micron. It is now possible to fabricate fibres of this thickness. They have relatively thick cladding to facilitate handling. One application of single mode fibres is to clean up the intensity profile of a light beam. If, for example, the intensity of the light beam measured in the plane perpendicular to the direction of propagation is required to be symmetrical and the existing beam has an irregular intensity pattern, injection of the beam into a single-mode fibre results in an emerging beam which has the intensity profile corresponding to that of the lowest mode.

12.3 CAVITIES

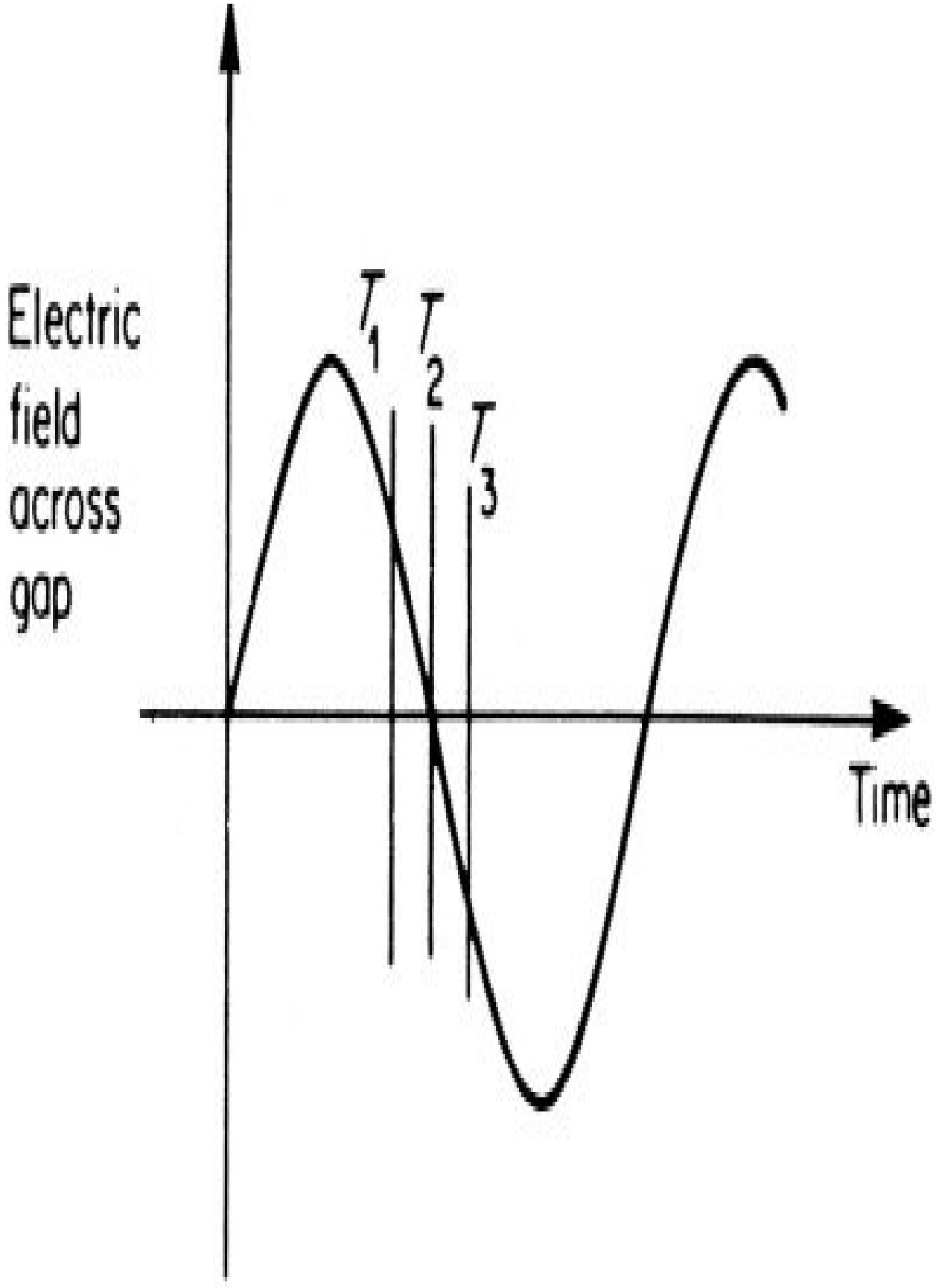
The oscillating electric currents that flow in antennas and give rise to radio waves, discussed in the next chapter, are produced in electrical circuits which use vacuum tubes and tuned lumped circuits or transmission lines. Vacuum tubes and the tuned circuits used at radio frequencies (up to 10^8 Hz) cannot be used at microwave frequencies (up to 10^{11} Hz). The devices usually used to cause the oscillations of charge which produce microwaves are called klystrons. These generators use resonating cavities in a somewhat analogous fashion to the way in which a low frequency oscillator uses a tuned LC circuit. The cavity stores large amounts of energy at the resonant frequency with very little energy loss per cycle of the fields. The Q of the cavity (see section 7.6) is thus very high, and the stored energy falls rapidly as the frequency goes off the resonant frequency. The generators thus oscillate at a resonant frequency of the cavity.

Figure 12.9. The basic structure of a reflex klystron. The electric field lines in the cavity go from side C to side D, and the field oscillates with time so that sometimes electrons are accelerated as they cross the gap, sometimes decelerated.

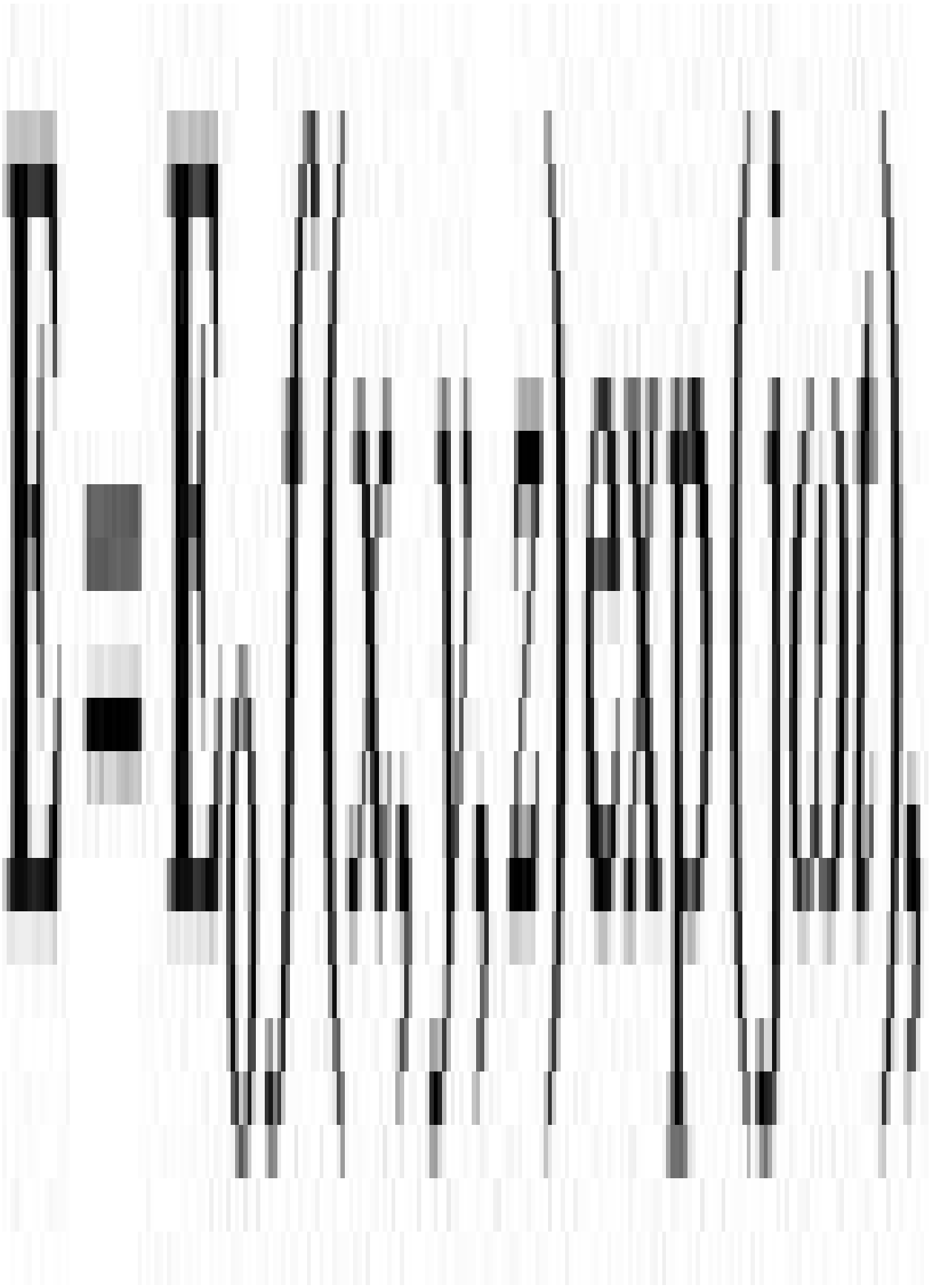

Outer casing of klystron tube: inside evacuated
Cavity resonator
Cathode
Heater
Reflector
C
D
Electron beam
Transmission line to waveguide
Heater supply
+V1
-V2

The simplest klystron, one frequently used in the teaching laboratory to generate waves of free space wavelengths a few cm at a power of a few hundred mW, is the reflex klystron. Figure 12.9 shows schematically the arrangement of electrodes and cavity resonator in such a device. These components are housed in an evacuated tube and the cavity is coupled to the waveguide either directly or via a short length of coaxial transmission line. Electrons from the heated cathode are accelerated to the gap in the cavity by the positive voltage V1. After exit from the cavity the electron beam travels towards a reflector electrode at a negative voltage −V2, and this causes the electrons to be repelled back towards the gap in the cavity. If we suppose that the klystron is oscillating, a high frequency electric field exists across the gap in the cavity. The electrons, on first passing through, thus undergo acceleration or deceleration, and their speed changes by an amount depending on the phase of the cycle of the electric field at the time the electrons entered the cavity. For example if they cross the gap during the time T1T2 in Figure 12.10 their speed is increased; if they cross during the time T2T3 their speed is decreased. Some of the electrons in the beam thus become bunched and the beam travelling towards the reflector no longer has a uniform density of electrons. For suitable combinations of the voltages V1 and V2 the electrons return to the gap, after being reflected, at a time when the phase of the field in the cavity is such that the field slows down the electron bunch. There is thus a net delivery of energy from the electrons to the cavity, which keeps the oscillations going. The frequency is tunable over a narrow range of about 10% by flexing the walls of the cavity resonator, since the resonant frequency of the cavity depends on its dimensions.

Figure 12.10. The electric field across the gap in the sides C and D of the reflex klystron plotted against time. The field is shown positive if the direction is such that electrons are accelerated when crossing the gap from left to right.

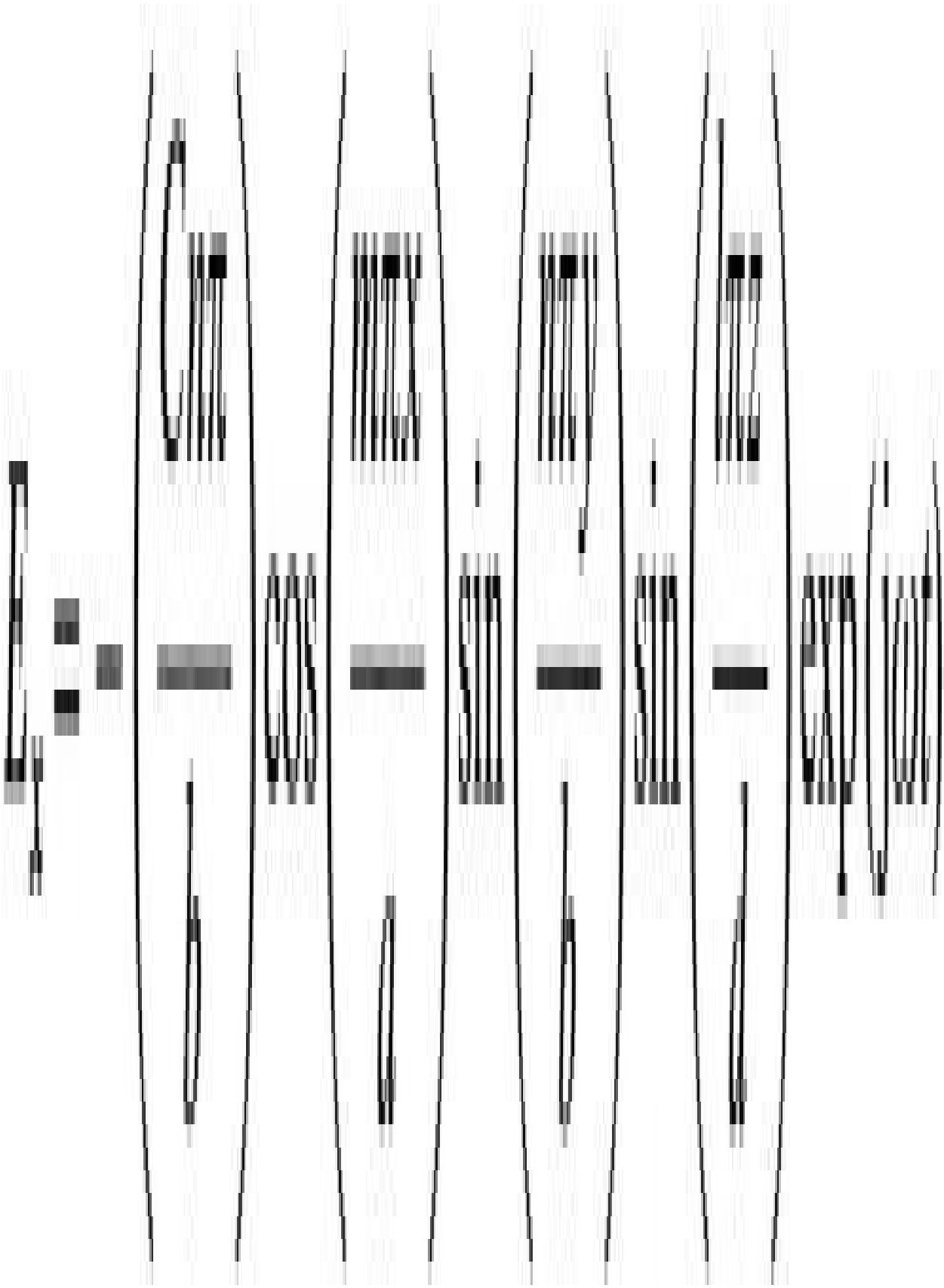

Electric
field
across
gap
T1
T2
T3
Time

The simplest cavity whose resonant frequencies can be calculated is a rectangular cavity. Consider the closed box shown in Figure 12.11 formed by adding two walls to the waveguide in Figure 12.4. The length of the cavity in the z-direction is d. The waves in the cavity are not progressive waves, they are standing waves. They may be written in the form

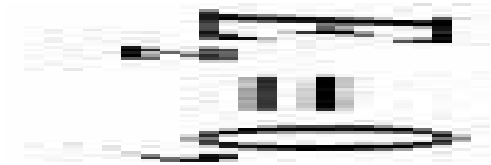


where ω is the angular frequency, and the function f(x, y, z) gives the spatial dependence of the amplitude of the wave. The form of this function can be guessed using our experience with the fields in waveguides. The field components Ex and Ey will now vary with the coordinate z, the variation being characterized by a third integer l. The electric field given by

(12.29)



(12.30)



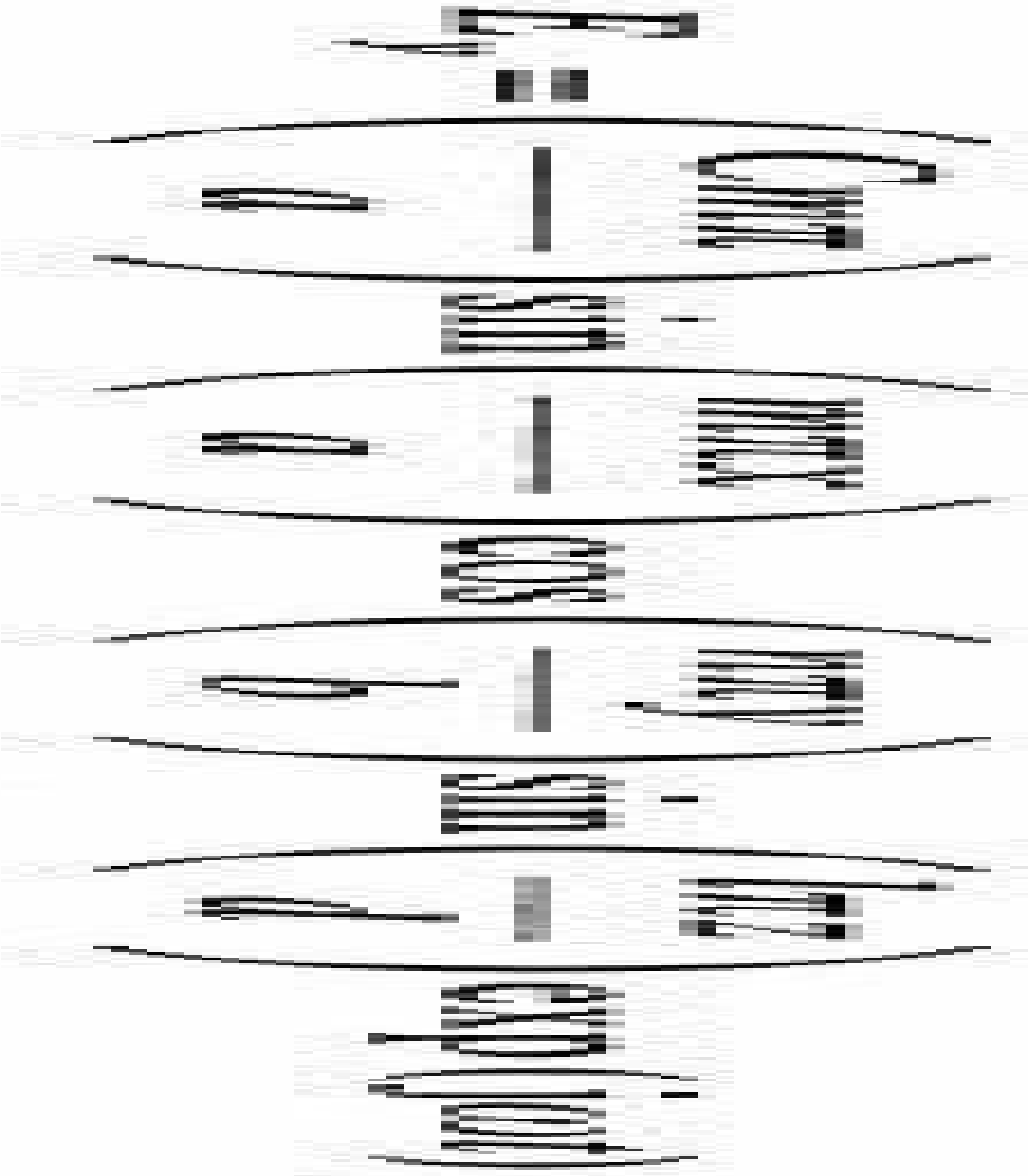
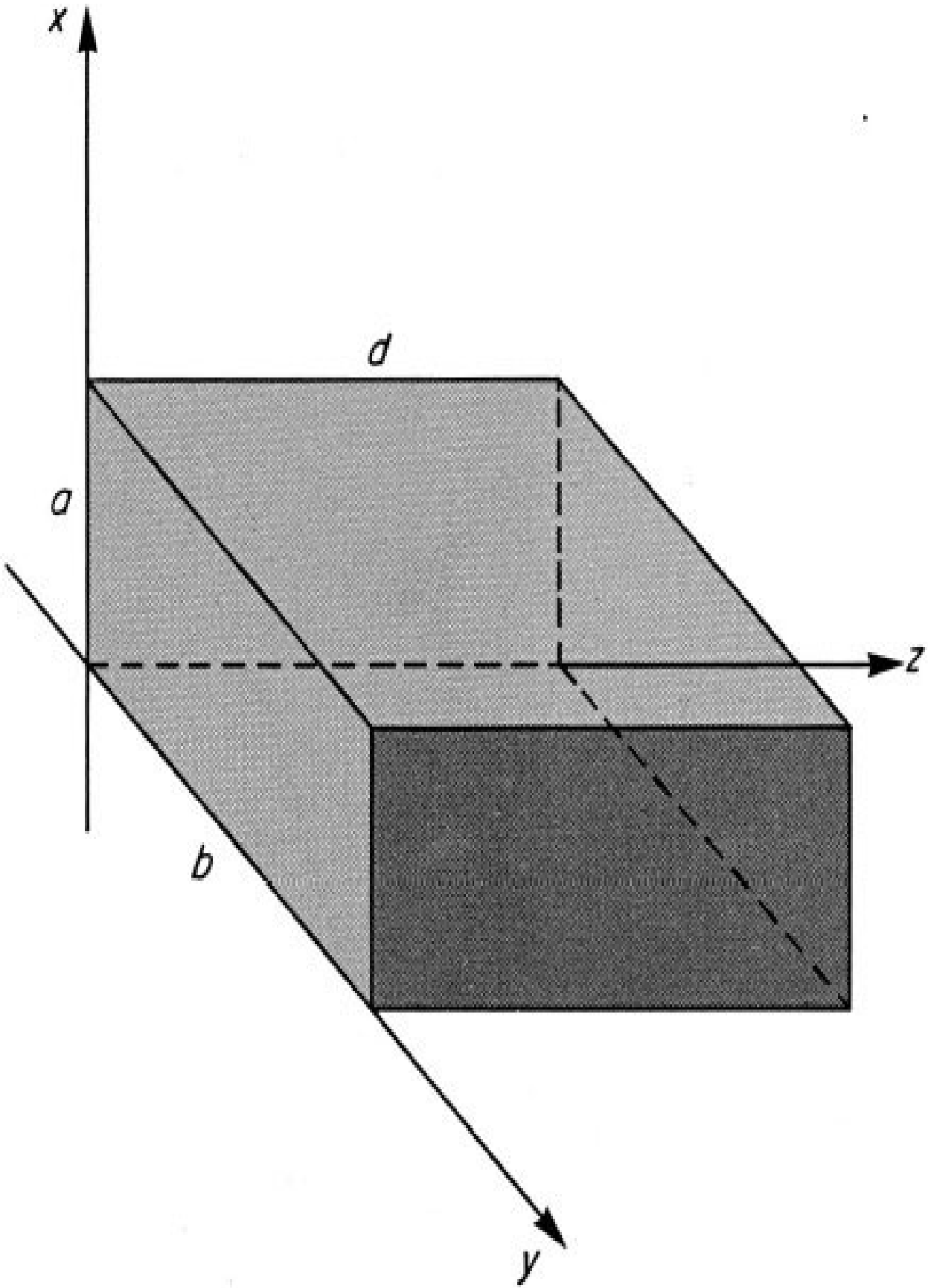
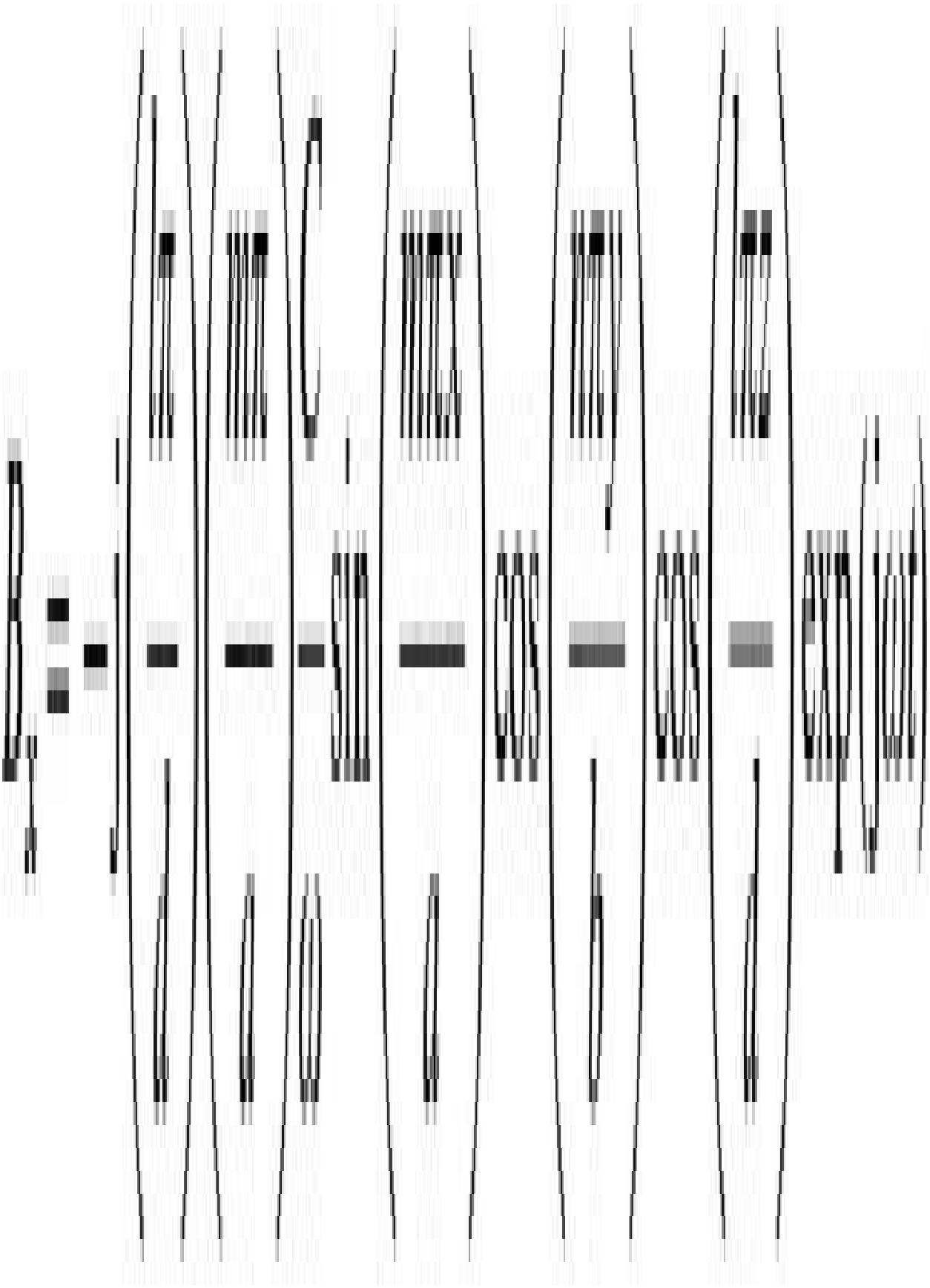


Figure 12.11. A rectangular cavity with sides of length a, b and d.

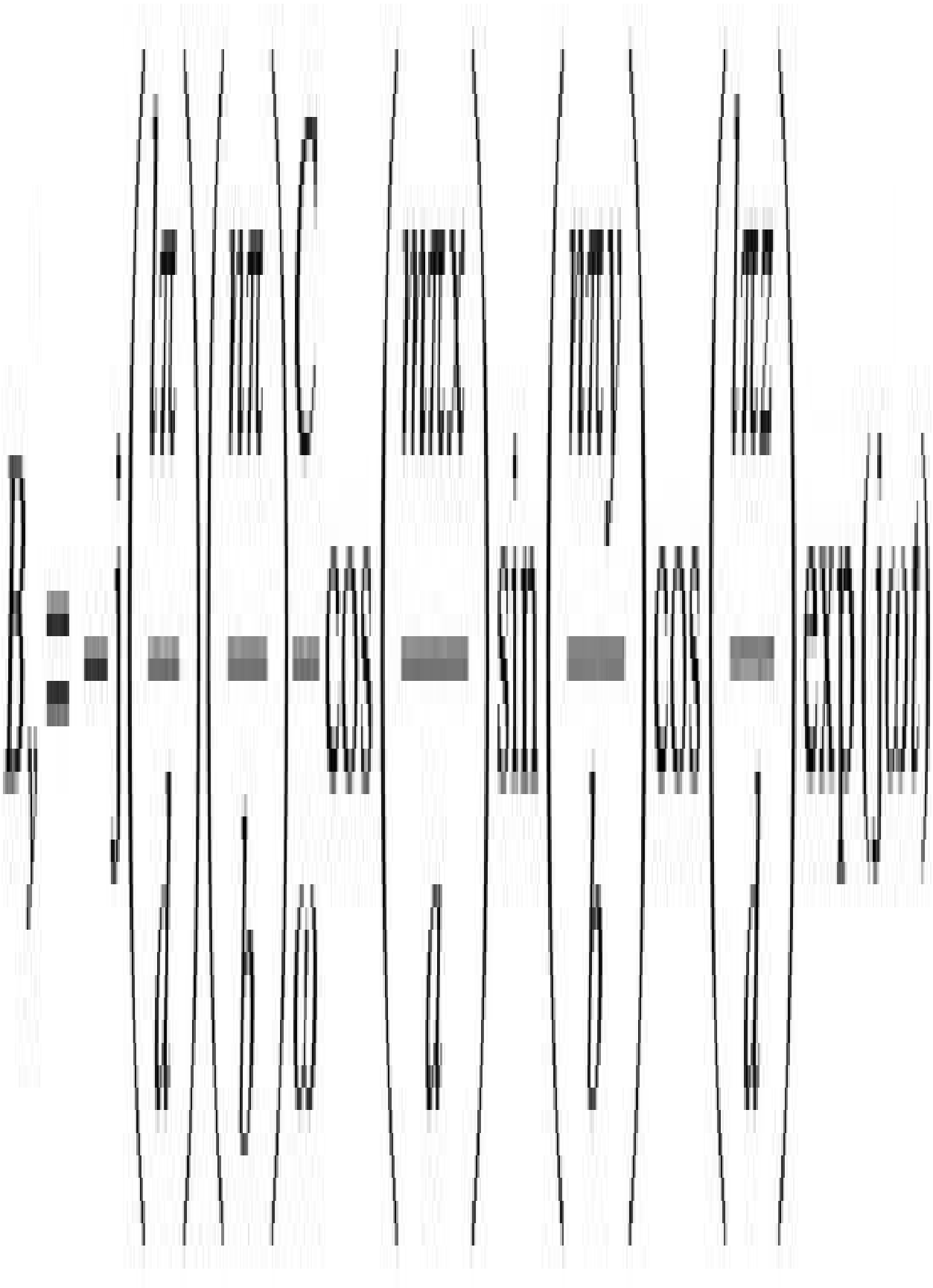

x
d
a
z
b
y

satisfies the boundary conditions on the cavity walls, and also satisfies the equation div E = 0. It can easily be shown using the Maxwell equation curl E = $-\partial B/\partial t$, or by extending the earlier solutions in an obvious way, that the magnetic field associated with this electric field is given by

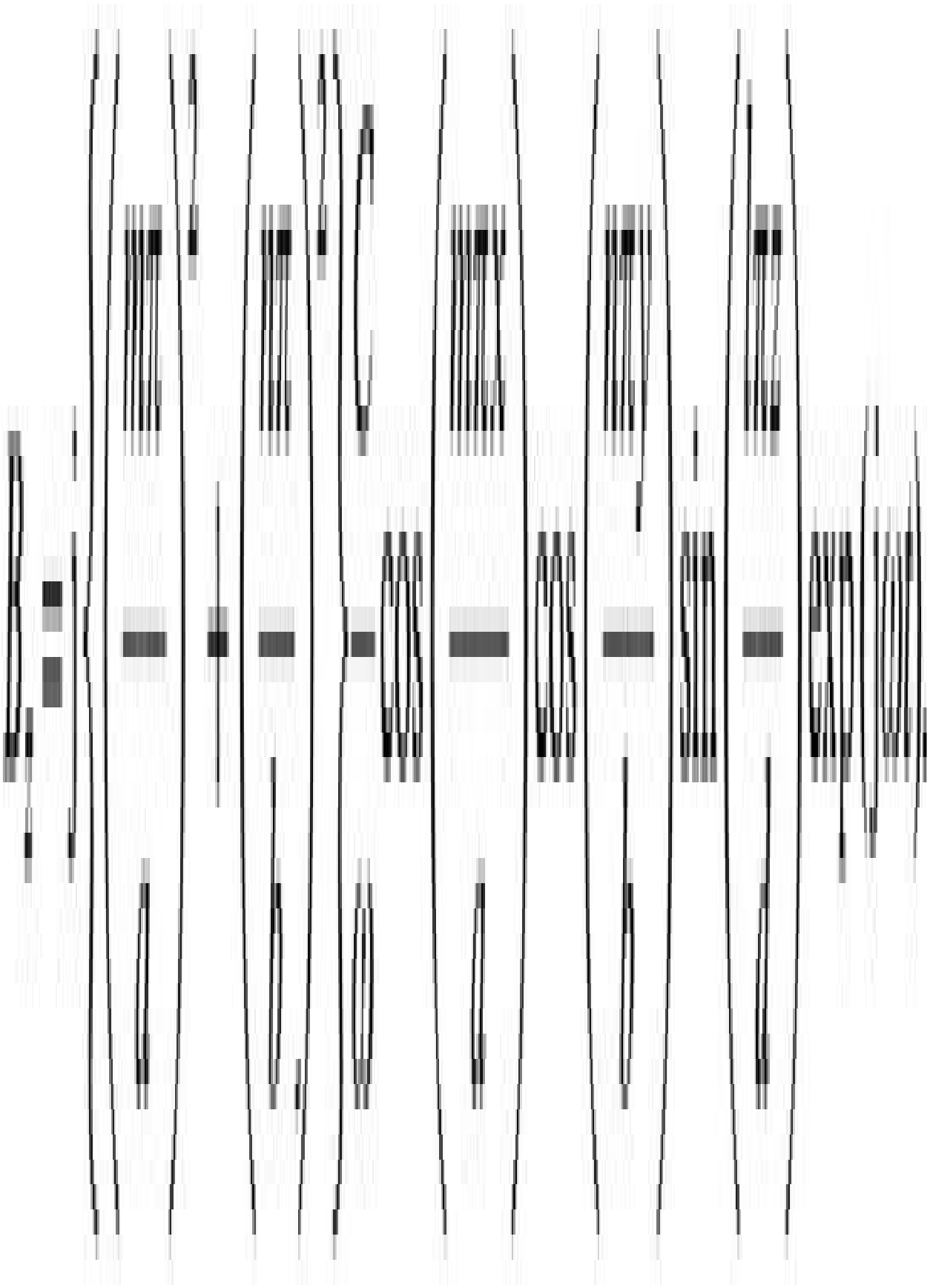
(12.31)



(12.32)



(12.33)



In order for these electric and magnetic fields to be acceptable solutions to Maxwell's equations inside the cavity, they must also satisfy the wave Equations (11.5) and (11.6). Substitution of any component into the appropriate wave equation shows that the fields are acceptable if the following equation is obeyed by the parameters m, n, l and k, the free space wavenumber of the radiation of angular frequency ω:

(12.34)

There is thus an infinite number of resonant frequencies of the cavity corresponding to different values of the integers m, n and l. The cavity is usually used so that a low mode is excited, e.g. the lowest frequency corresponds to two of the integers equal to one, and the third integer, that associated with the shortest side of the cavity, equal to zero. Since the Q values of the cavity at each resonance are very high, typically a few thousand, when the lower resonances are used there is no overlapping.

The m = 0, n = 1, l = 1 resonance in which the fields are obtained by the appropriate substitutions in Equations (12.29) to (12.33) is designated the TE011 resonance because the electric field has no z-component. There is one other possible set of solutions for the electromagnetic fields in the cavity. These are TMmnl standing waves in which the magnetic field has no z-component, but the electric field has a non-zero z-component. These waves occur at the same frequencies as the TEmnl waves (for the same cavity), and so at a particular resonance the standing wave is the sum of two resonating waves, the TE mode and the TM mode*.

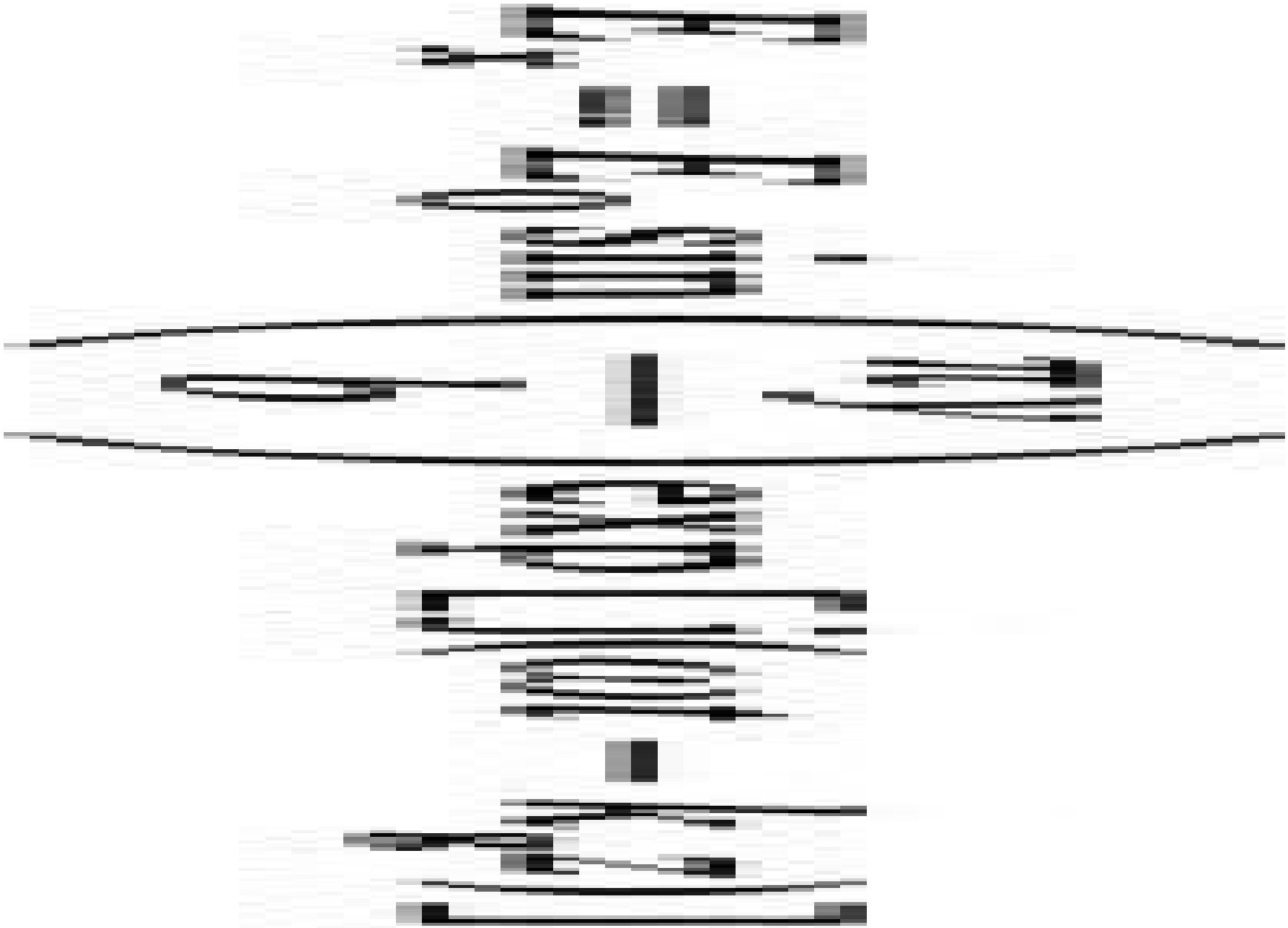
Cavities are often used as frequency meters as well as in microwave generators. A cavity will readily absorb energy from a waveguide (usually via a small hole in one wall) when the frequency of the wave is near a resonant frequency of the

cavity. To measure frequency, a plunger is used to vary the length of the longest side of the cavity, which is operated near the TE011(TM011) resonance. An absorption dip in the energy in the guide at points beyond the cavity indicates the resonance, and the length of the cavity side then gives the frequency from Equation (12.34). The resonant frequencies of cylindrical cavities are also amenable to calculation, and since cylindrical cavities can be made more accurately than rectangular cavities they are more often used in very accurate frequency measurements. Another use of cavities is in applications where high microwave field levels are required. For example in electron spin resonance experiments† in which a sample is placed in a high frequency magnetic field of large amplitude, the centre of a rectangular cavity resonating in the TE021 mode is often used.

Cavities are also used to accelerate charged particles to high speeds in nuclear accelerators. The way they work as accelerators is in many ways the reverse of the mode of operation of the klystron, discussed in section 12.2.2. In the klystron, bunches of electrons given energy by an external steady accelerating voltage, arrive at a cavity at times such that they feed energy into the oscillating field, thereby keeping the oscillations going. In an accelerating cavity the oscillations are maintained by an external source of radiofrequency power. If bunches of charged particles cross the cavity during parts of the field cycle when the average field on them is in their direction of motion, they are accelerated, acquiring energy from the external radiofrequency power source. The energy drain on the source which goes into kinetic energy of the beam is only a small fraction of the total power delivered by the source. Most power is dissipated in the cavity wall. For example, if an accelerator produces a beam of 100 μA of protons of kinetic energy 50 MeV the energy given to the particles per second is 4.8 kW. The rate of energy dissipation in the walls of the cavities used to perform the acceleration would typically be about a hundred times this. This energy drain in the walls can be eliminated by using cavities with superconducting walls, although some energy is still needed to keep the surfaces at a sufficiently low temperature that they remain superconducting. There are now several superconducting cavity accelerators in operation at various sites worldwide, each with complex cavity designs tailored to the particular accelerator function.

PROBLEMS 12

12.1 Show that the field component



corresponds to one of the TE waves propagated in the z-direction in a rectangular waveguide whose width in the y-direction is b. Show that the component Ey in the wave is zero. Determine the cut-off frequency of the wave.

12.2 Determine the maximum and minimum widths of a waveguide of square cross-section if it is to transmit waves of free-space wavelength λ in the TE01 mode only.

12.3 What is the maximum average power which an air-filled guide can transmit in the TE01 mode if the dimensions of the guide are a = 1 cm, b = 2 cm, and the frequency is 10^{10} Hz? The breakdown electric field in air is 30 kV cm−1.

12.4 A source maintains electric fields of constant amplitude in the TE01 mode in a rectangular waveguide. Discuss how the power received at the far end of the guide varies as the frequency of the radiation is reduced. If the guide is terminated with an impedance such that there is no reflected wave what is the energy density in the waveguide?

12.5 Calculate how the wave and group velocities of the TE01 wave in a rectangular guide with a = 1 cm, b = 2 cm vary with frequency.

12.6 Using the method outlined on page 423, estimate the attenuation of the TE01 mode in a brass rectangular waveguide with a = 1 cm, b = 2 cm at a frequency of 10^{10} Hz. The conductivity of brass is 1.6×10^{7} $(\Omega m)^{-1}$.

12.7 A rectangular cavity has dimensions a = 2 cm, b = 3cm, d = 4 cm. How many resonances are there within the frequency range $v = 5 \times 10^{9}$ Hz to $v = 10^{10}$ Hz?

12.8 Show that the assumption that a TEM wave can be propagated within the interior of a hollow metal pipe leads to contradictions with electromagnetic theory. This constitutes a proof that there can be no TEM waves within a waveguide.

12.9 A crystal diode is used in a slotted waveguide to measure the power radiated by a horn aerial at the end of the guide. The maximum and minimum values of the voltage output from the diode detector as it is moved along the slotted guide are in the ratio 9 to 1. What fraction of the power in the wave

travelling down the guide is radiated by the horn?

12.10 A rectangular cavity with dimensions a, a, 2a, in the x, y, and z directions respectively, has walls of very high conductivity and is excited in a mode for which the electric field components are

$$E_y = E_0 \sin(\omega t) \sin(\pi z/a) \exp(j\omega t)$$

$$E_y = E_x = 0$$

What are the components of the magnetic field B, and what is the angular frequency ω of the resonance?

*For example the TEM waves inside a coaxial cable transmission line have the electric field radially out from the centre wire, whilst the magnetic field lines are circular around the centre conductor.

*In the theory of black body radiation (see for example, F. Mandl, Statistical Physics, 2nd edition, J. Wiley, London, 1988, Chapter 10) one needs to calculate the number of independent resonating modes of a cavity within a given frequency interval. At each resonance there are two independent waves, the TE and the TM, and this gives rise to a factor of two in the density of states function.

†See for example, C. Kittel, Introduction to Solid State Physics, 6th edition, J. Wiley, New York, 1986, Chapter 16.

CHAPTER 13

The generation of electromagnetic waves

In this chapter we discuss the electromagnetic fields produced by specified changing charge and current distributions. Radio waves are produced by currents oscillating at radio frequencies in aerials; microwaves with wavelengths of a few centimetres can be produced by making electrons oscillate to and fro at frequencies of about 10^{10} Hz in a tuned cavity; visible light arises from very rapid readjustments of the charge and current distributions within the electron clouds of atoms. The emission of visible light can only be fully described by quantum mechanics. However the classical radiation theory we will outline is a necessary guide to the quantum theory of radiation.

Any distribution of changing charge and current acts as a source of electromagnetic radiation. Far away from a source the electric and magnetic fields fall off inversely proportionally to the distance from the source. Nearer to the source there are other components of the fields which fall off more rapidly with distance, but these components make no contribution to the power continuously radiated. The part of the field inversely proportional to distance is called the radiation field, and the region far away from the source where the radiation field is the only important one is called the radiation zone. In Chapter 11 we considered the plane wave approximation to radiation fields. At sufficiently great distances from any source the radiation field can be adequately represented for most purposes by plane waves. In this chapter we derive expressions which enable us to describe more completely the fields of a given system of changing currents or charges. We then use the expressions to determine the total power radiated and the distribution of the radiated energy over space in a few simple examples.

13.1 THE RETARDED POTENTIALS

In the discussions of Chapters 11 and 12 on electromagnetic waves in free space, or in media where there were no free charges or currents present (other than those induced by the wave itself), there was no need to introduce the potentials A and ϕ. We were able to solve the problems quite simply without them. However if we wish to relate the fields to the changing charge and current distributions which produce them it is far easier to work with the potentials than with the fields directly.

In this section we will outline the steps taken to obtain the potentials in terms of given charge and current distributions. Once the potentials have been determined, the fields can be obtained from the relations already given in Equation (4.49) and (6.18), viz.

(13.1)

and

(13.2)

We will first obtain differential equations obeyed by the potentials A and ϕ separately. We will only consider changing charge and current distributions in free space (or air), so that Maxwell's equations become:

(13.3)

(13.4)

(13.5)

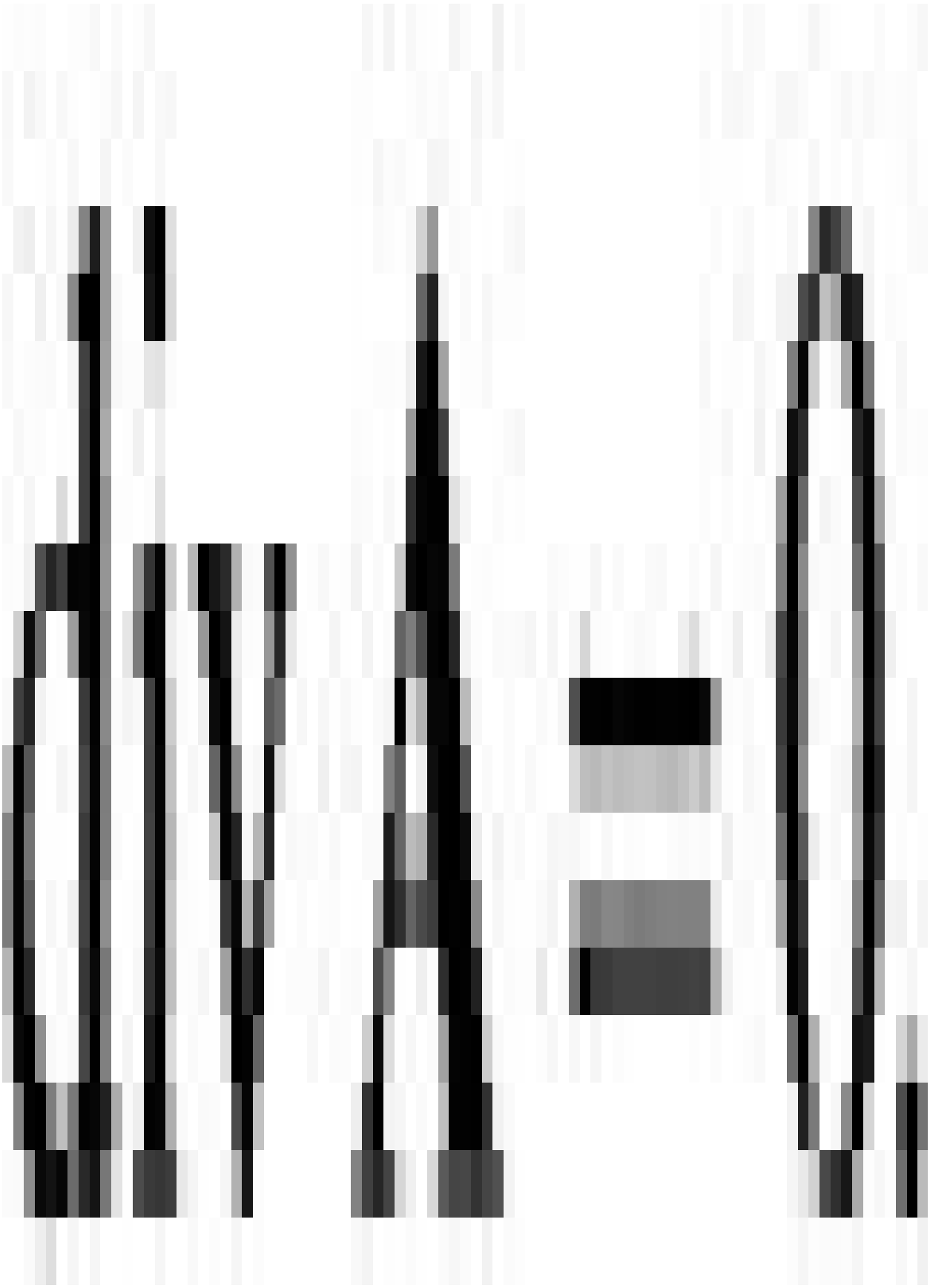
(13.6)

Substitution of the expression (13.2) for the field E into Equation (13.3) gives

(13.7)

The potentials A and ϕ are not uniquely specified by Equations (13.1) and (13.2). We can add to the vector A any function whose curl is zero and the same magnetic field results: we can add to the potential ϕ any function whose gradient is zero and the same electric field results. We are at liberty to impose any condition on the functions A and ϕ that does not contradict Maxwell's equations.

In the discussion of steady magnetic fields we added the condition



In dealing with time-dependent fields it is convenient to generalize this to

(13.8)

The restriction on the potentials imposed by Equation (13.8) is known as the Lorentz condition. It involves no contradictions or difficulties; we can choose potentials A and ϕ which with Equations (13.1) and (13.2) give specified fields B and E everywhere and satisfy Equation (13.8). Different potential functions A and ϕ which also give the specified fields can be obtained by requiring that different relationships between A and ϕ be satisfied. However, imposing the Lorentz condition (13.8) introduces a symmetry into the equations for the potentials that is very useful and elegant. It also enables the formalism of electromagnetism to be brought into line with the special theory of relativity as discussed in the next chapter.

If we substitute Equation (13.8) into Equation (13.7) we obtain

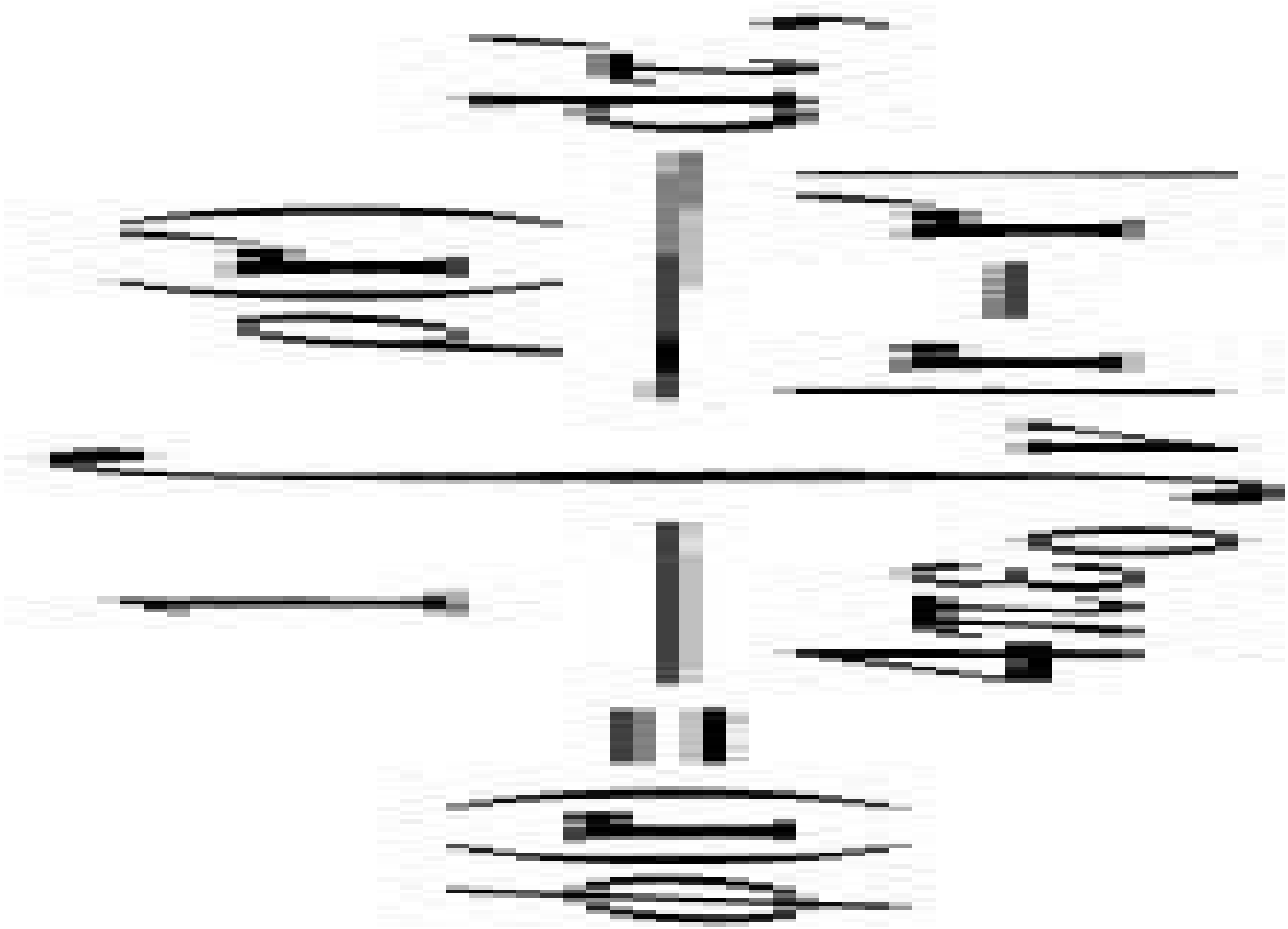
(13.9)

an equation involving ϕ alone. It may be noted that in those regions of space where the charge density ρ is zero, the potential ϕ obeys the same wave equation as the fields E and B. The equation for the vector potential A can be obtained by substituting expressions (13.1) and (13.2) for the electric and magnetic fields into Equation (13.6). The use of the Lorentz condition leads to an equation similar in form to Equation (13.9):

(13.10)

We have now obtained differential equations for the potentials and it remains to solve them to give expressions for A and ϕ in terms of the charge and current distributions. We will not solve the equations formally* but will introduce the solutions by analogy with the solutions to the time independent problems discussed in electrostatics and magnetostatics. The solution to an electrostatic problem is obtained by solving the equation

The solution has already been given in section 3.1, and is



where V is the volume occupied by the charge distribution ρ. Now consider the situation where ρ changes with time. Let the point P on Figure 13.1 have position vector r, and the point Q have position vector r′. If the charge distribution near Q changes with time, the information that it has changed can only be appreciated at the point P a time $|r - r'|/c$ after the change. This is the time it takes electromagnetic radiation to travel the distance $|r - r'|$. Hence it is plausible that at a time t the contribution to the potential $\phi(r, t)$ at P from the charge within the volume element $d\tau'$ at Q depends not on what the charge is at time t but on what it was at the time $t - |r - r'|/c$. This is true for all the volume elements within the volume V. (Note that the time lapse $|r - r'|/c$ depends on the position vector r′ and so varies over the volume V.) The required solution to Equation (13.9) is thus obtained by integrating the contributions over the volume V,

(13.11)

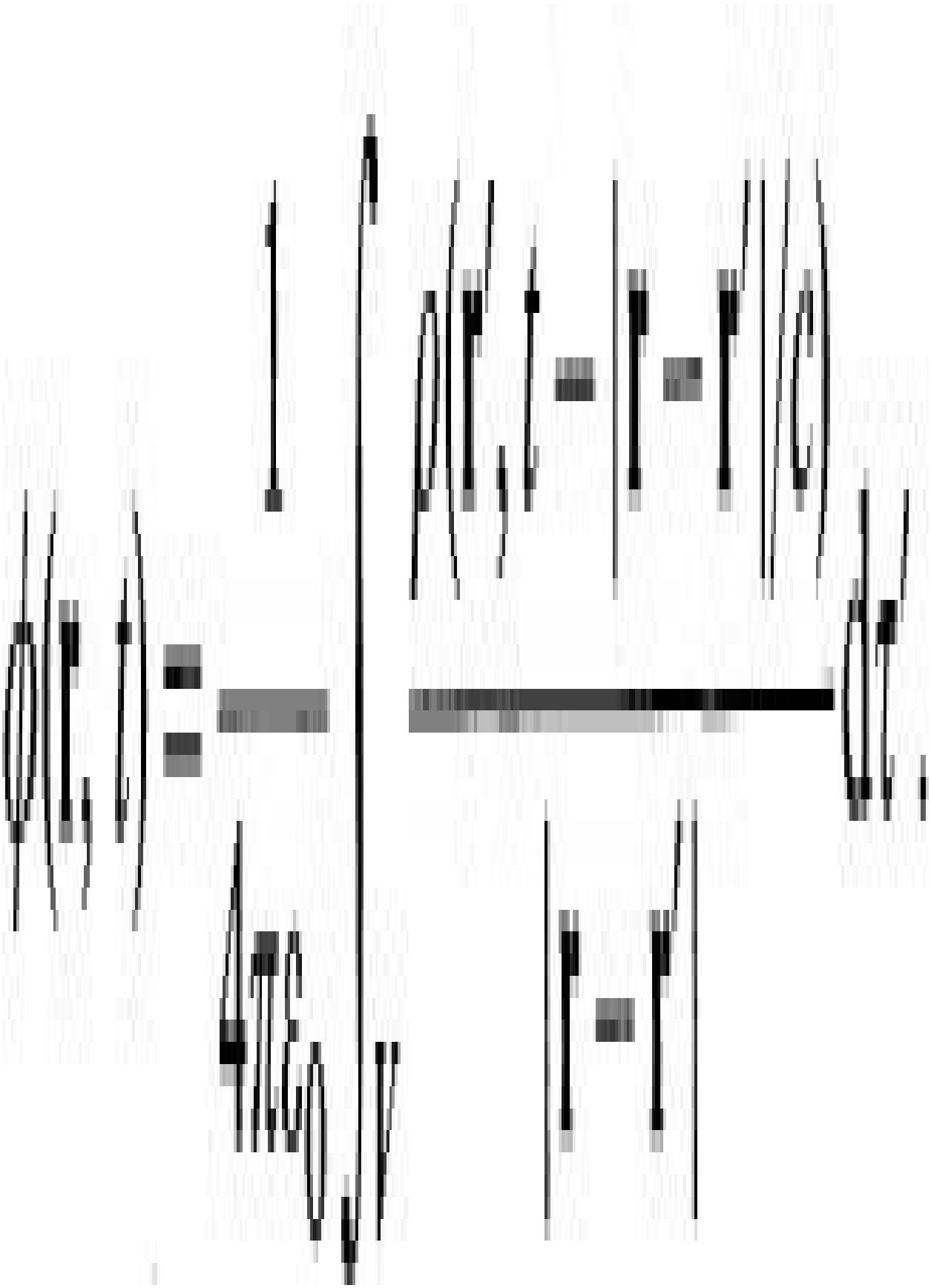
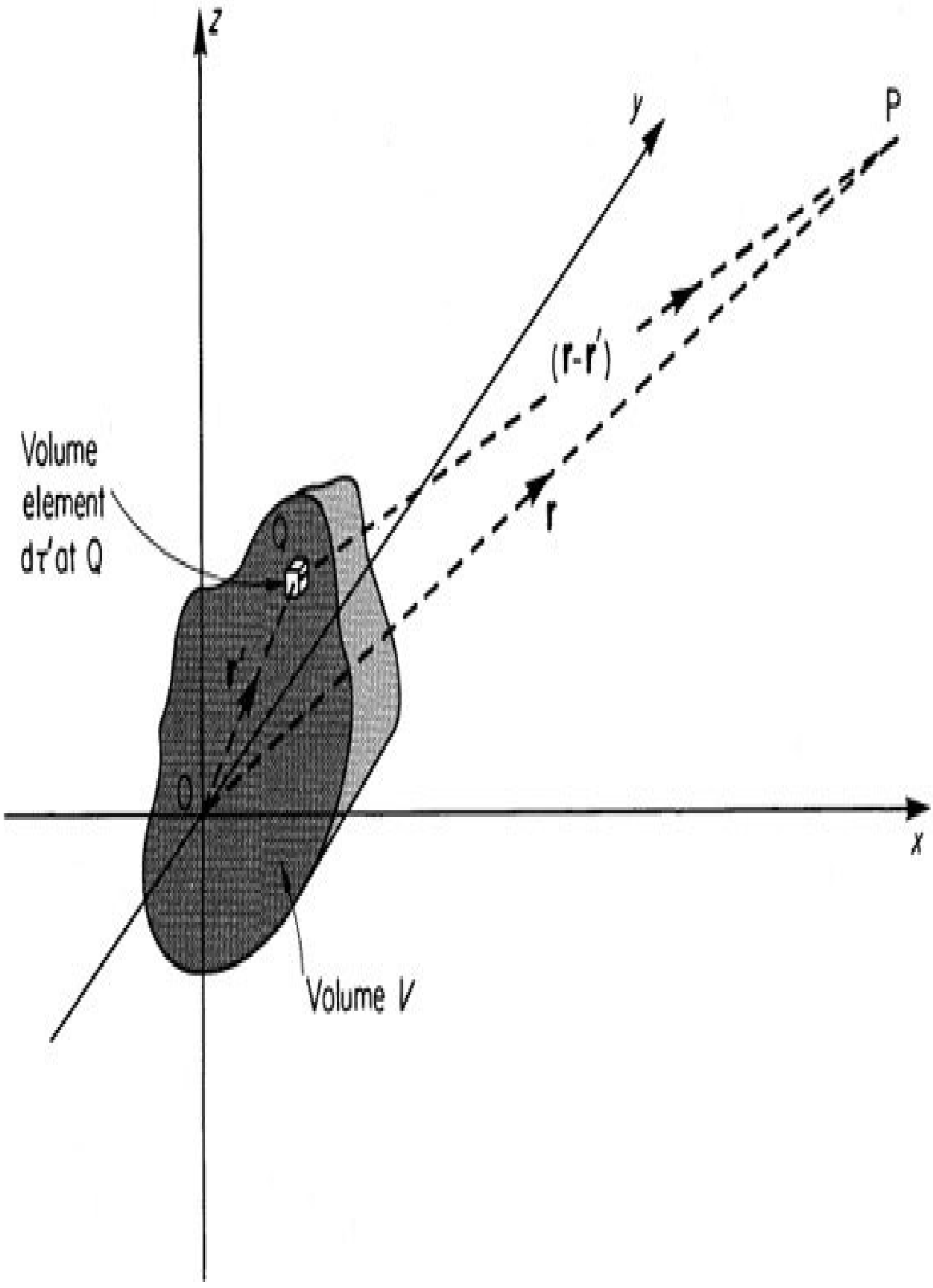
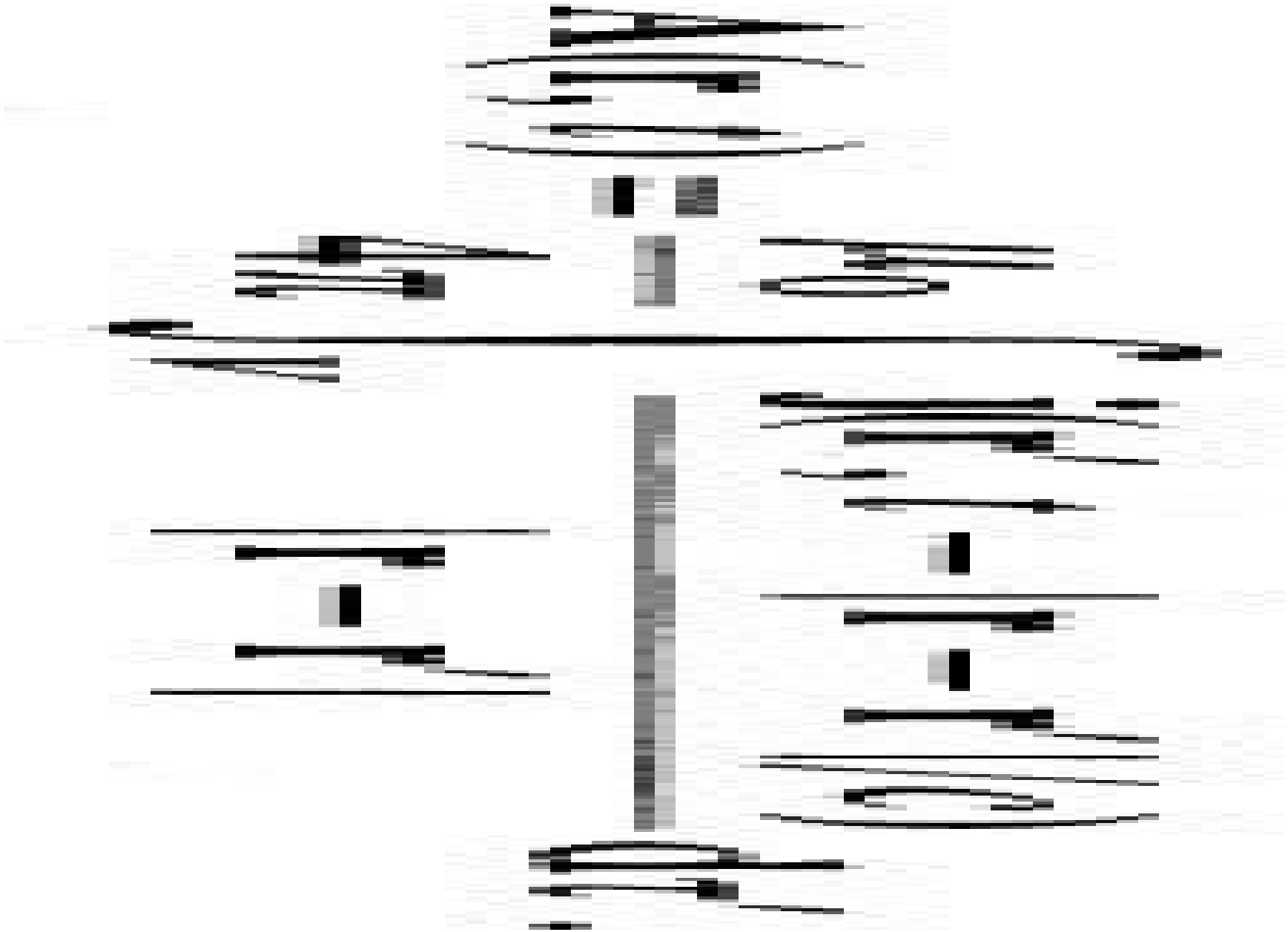


Figure 13.1. The coordinates used in the discussion of retarded potentials.


z
y
P
(r-r')
r
Volume
element
dτ' at Q
Q
r'
O
x
Volume V

The solution to Equation (13.10) is obtained in a similar manner by beginning with an equation which gives the magnetic vector potential due to a steady distribution of current (see Appendix C). If the current distribution within a volume V is changing then the potential A is made up of contributions which depend on the magnitude of the current density at appropriate times before the instant of observation. The solution to Equation (13.10) is

(13.12)

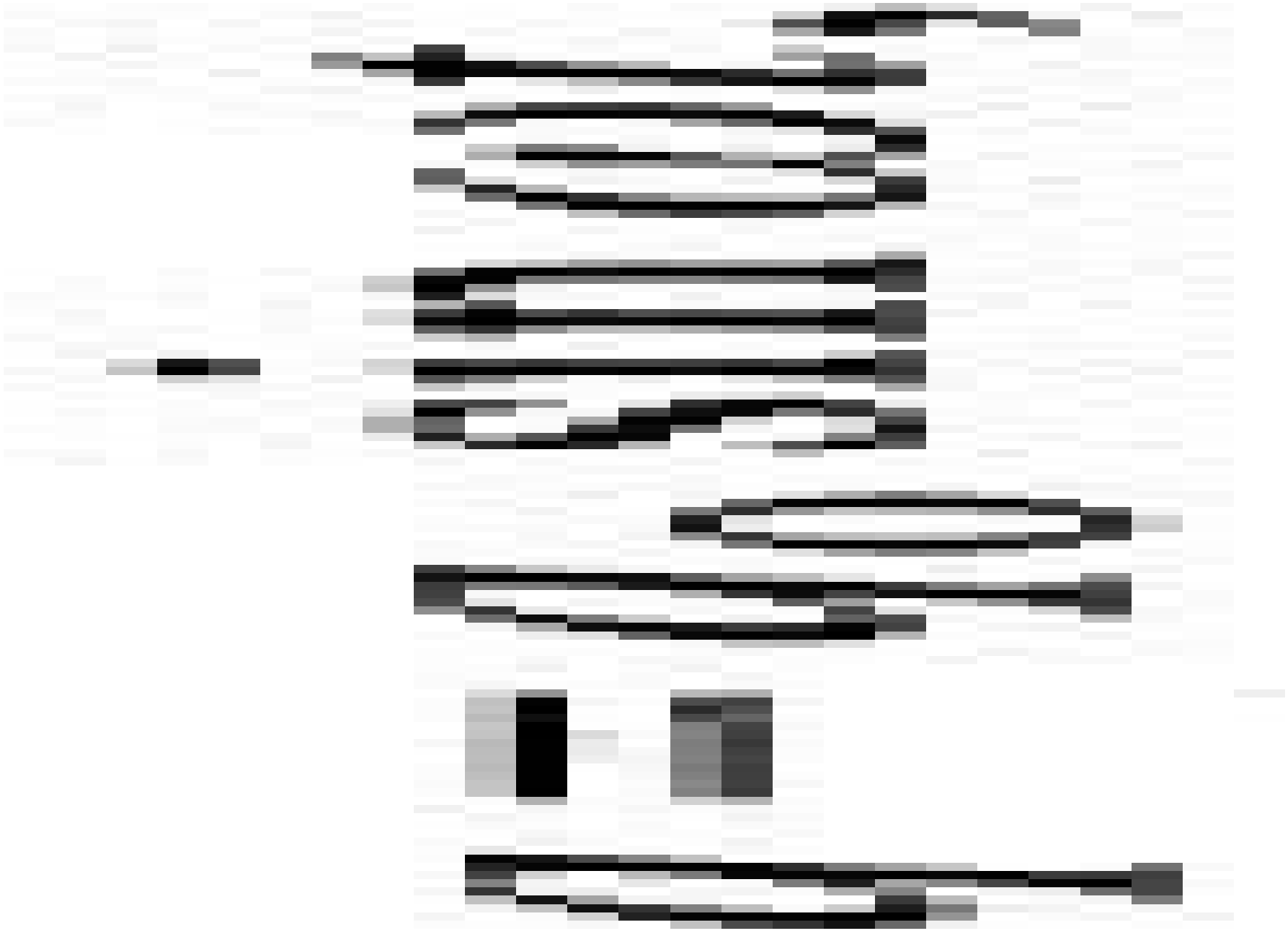


The potentials A and ϕ given by Equations (13.12) and (13.11) are known as the retarded potentials. The solution to the time dependent problem is just like the solution to the static problem, except that one sees the different parts of the source as they were at earlier times, the times taken by light to travel the apppropriate distances.

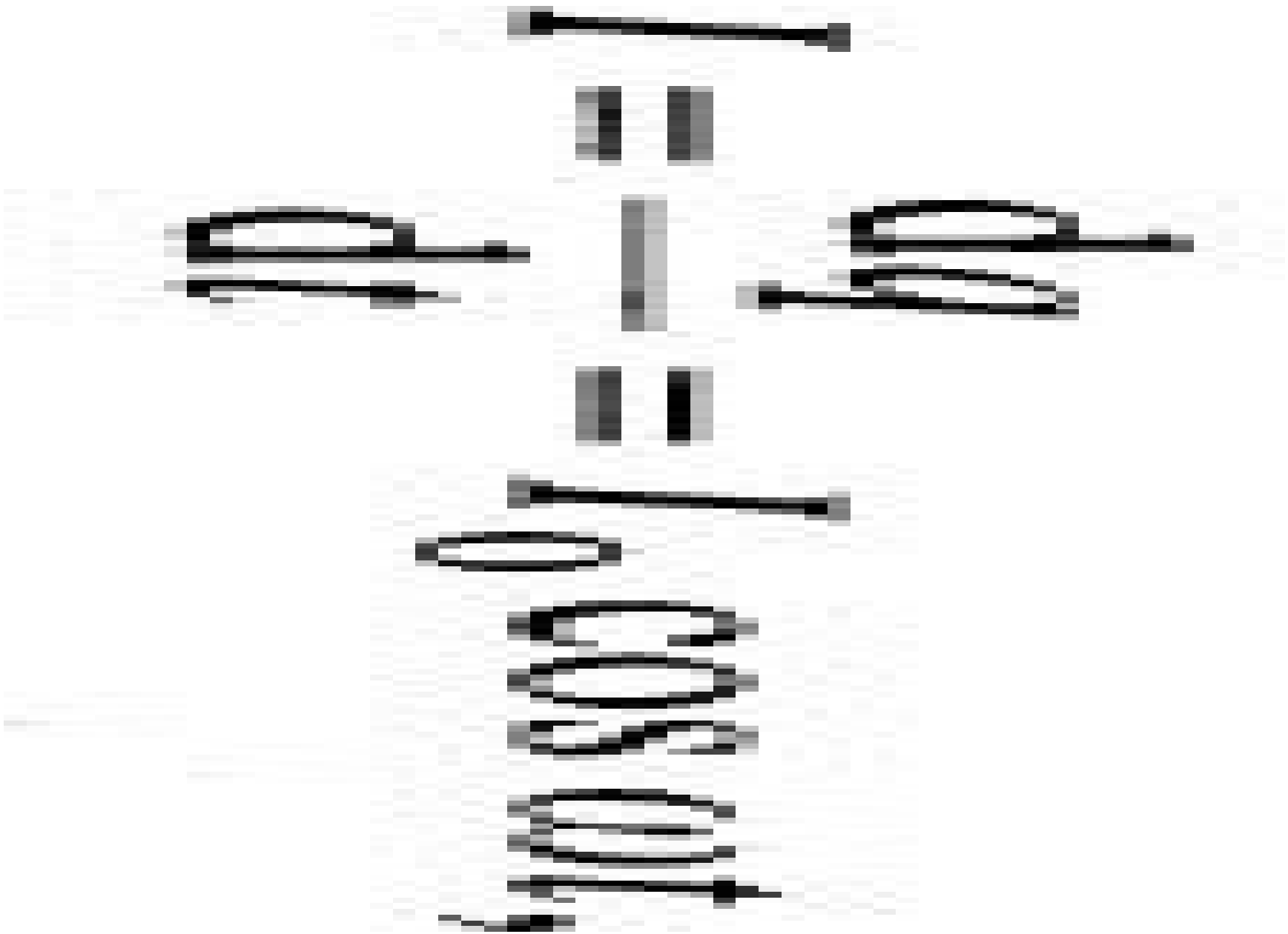
It is usually no easy matter to obtain the retarded potentials from given charge and current distributions, or to calculate the fields E and B using Equations (13.1) and (13.2) if the potentials have been found. Complications arise due to the fact that the retarded times are different for different parts of the source. In certain simple situations, and with sensible approximations, the fields can be determined. An example is given in the next section.

13.2 THE HERTZIAN DIPOLE

In this section we calculate the radiation fields in the simple but important example of an oscillating electric dipole. Suppose that charge is transferred periodically from a very small sphere at one end of a short wire to another small sphere at the other end. If the charge on one sphere at time t is q, the charge on the other sphere is – q, and if the variation of the charge with time is sinusoidal q is given by



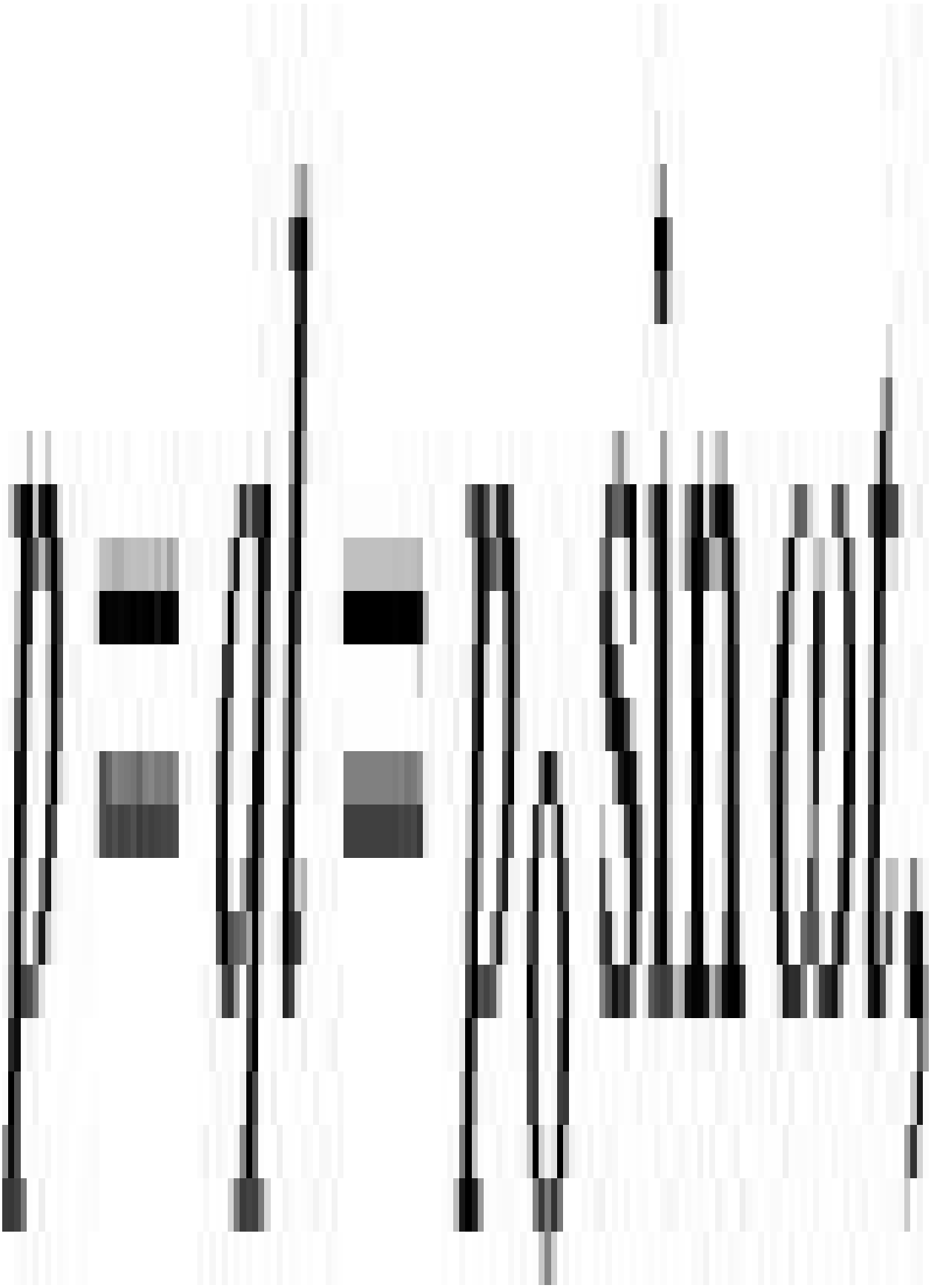
where q_0 is the amplitude of the oscillating charge and ω is the angular frequency of the oscillations. If the length of the wire is very short compared to the wavelength of the radiation produced, the current I in the wire can be taken to be the same at all points along its length, and



where

(13.13)

The charge oscillating between the two spheres is equivalent to an oscillating dipole moment. If the distance between the spheres is l the instantaneous dipole moment P is given by



where

(13.14)

The small oscillating dipole is called a Hertzian dipole. The radiation field of such a dipole is important in the quantum theory of the emission of radiation by atoms, molecules and nuclei. A readjustment of the charge distribution within the microscopic system plays the role of the oscillating dipole, although in individual atoms or molecules the change in the charge distribution is not continuous, and is associated with a single photon rather than a steady emission of energy.

We will now determine the total power radiated by a Hertzian dipole, and the way in which the radiation field is distributed throughout space. We follow the procedure outlined in section 13.1: firstly the retarded potentials are calculated, making assumptions appropriate to the small dipole; secondly the electric and magnetic fields are calculated from the potentials; finally the Poynting vector gives the energy radiated. The dipole is shown in Figure 13.2. We have chosen the z-axis to lie along the wire, and the origin of coordinates to be at the centre of the wire. We will first calculate the vector potential A. Let us suppose that the wire is very thin, so that when the potential is calculated from Equation (13.12) the integral over the volume occupied by the current becomes a line integral over the length of the wire. The vector $j(r', t - |r - r'|/c)d\tau'$ can then be replaced by the vector $I(r', t - |r - r'|/c)dI'$, where dI' is a vector in the direction of the wire at the point with position vector r'. In the present example the direction of the current is always along the z-axis, and the vector $I(r', t - |r - r'|/c)dI'$ becomes a vector of magnitude $I(z', t - |r - ezz'|/c)dz'$ with direction along the z-axis. Hence the magnetic vector potential is everywhere parallel to the z-axis, the components Ax and Ay always being zero. The component Az is given by Equation (13.12),

(13.15)

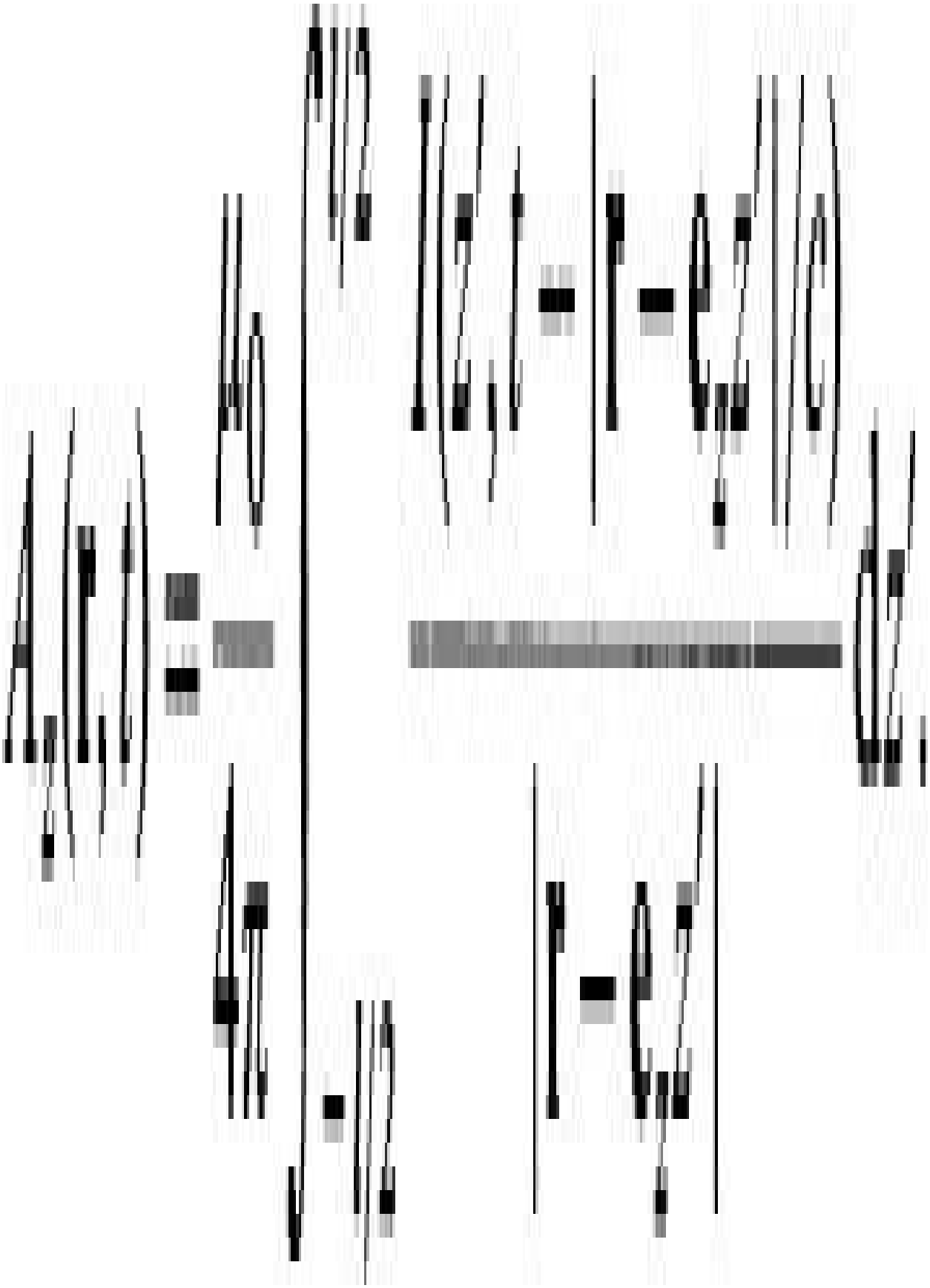
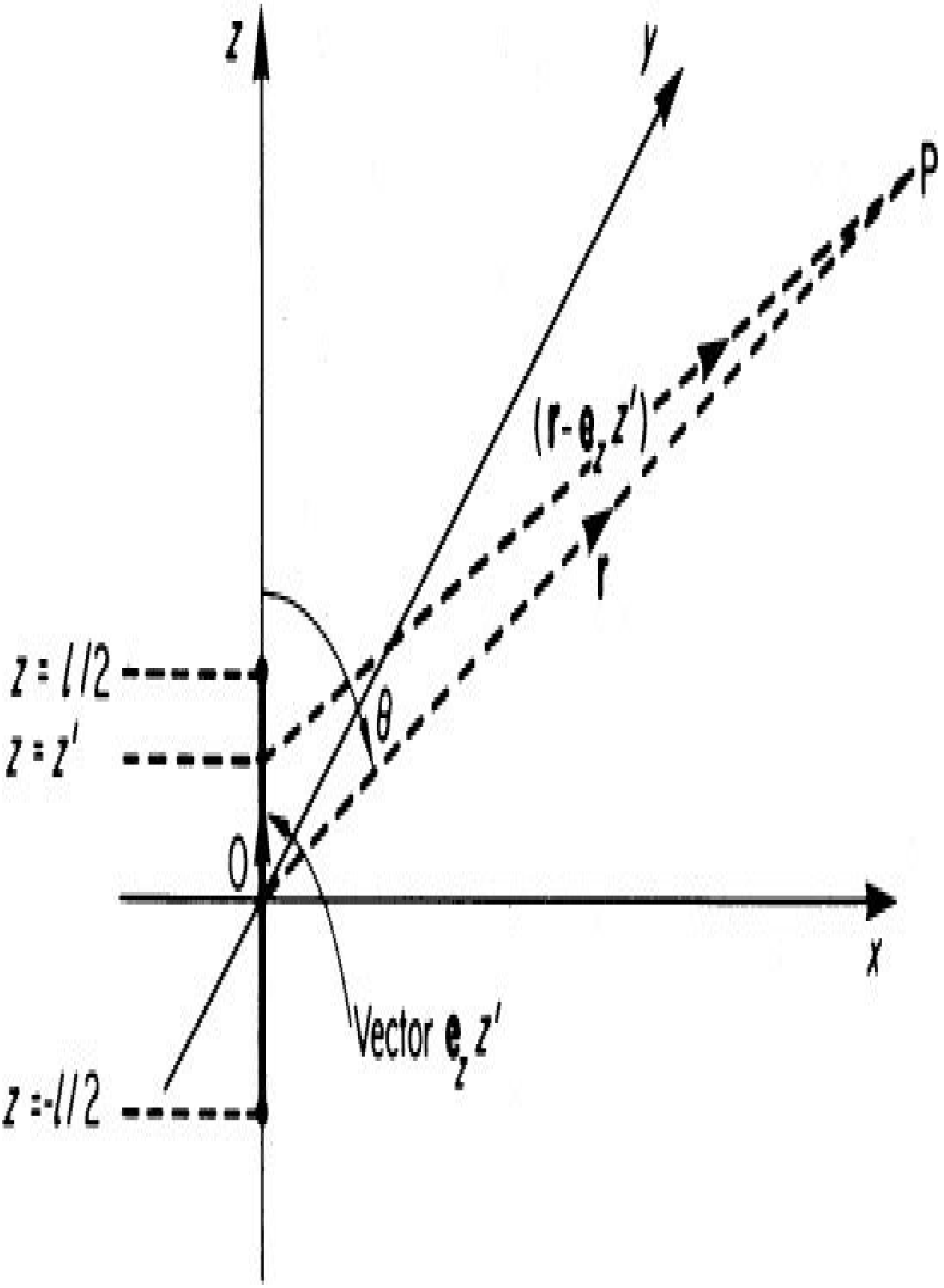
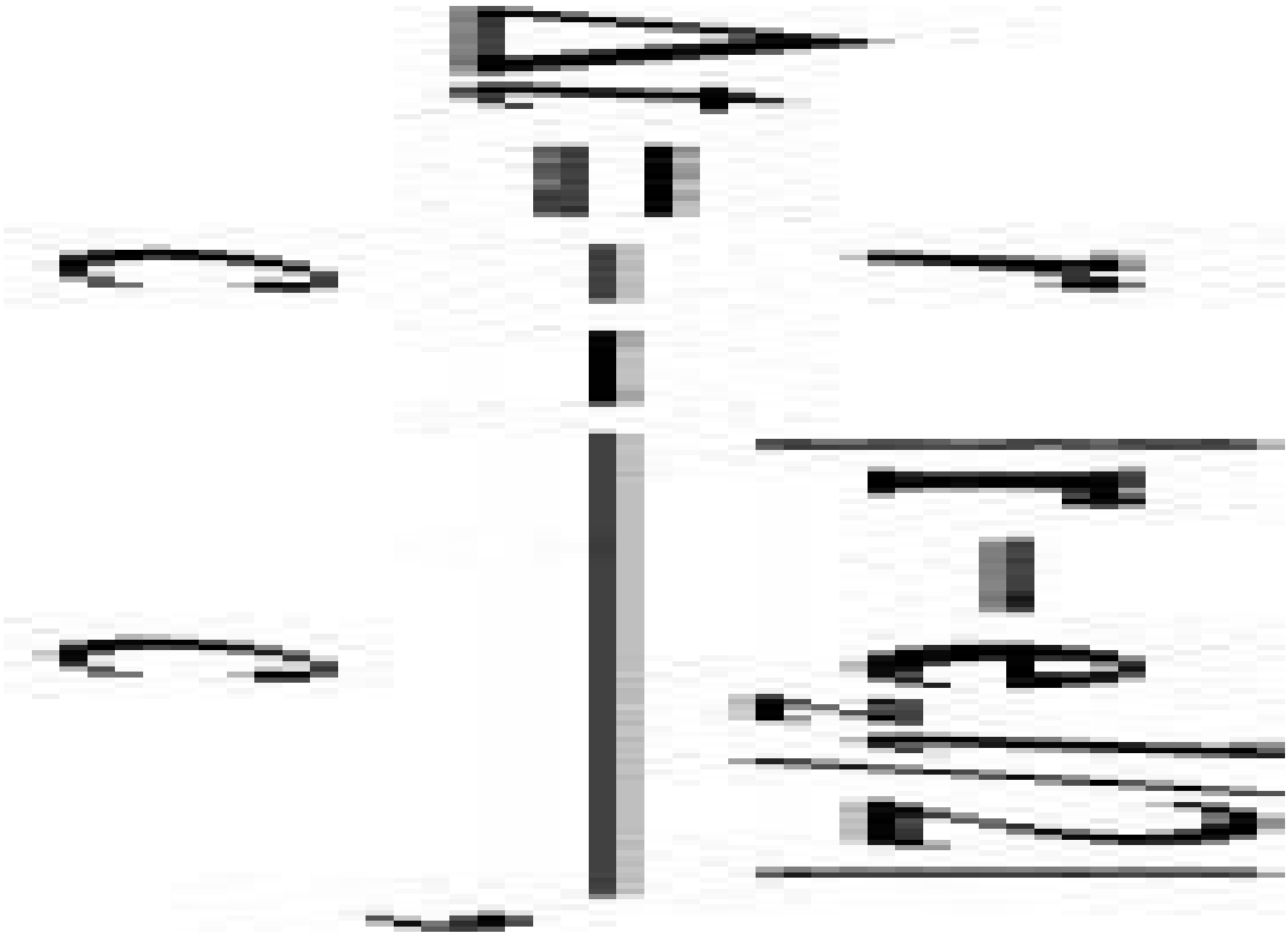


Figure 13.2. The coordinates used in the discussion of the Hertzian dipole.


z
y
P
(r - e_z z')
r
z = l/2
z = z'
θ
O
x
Vector e_z z'
z = -l/2

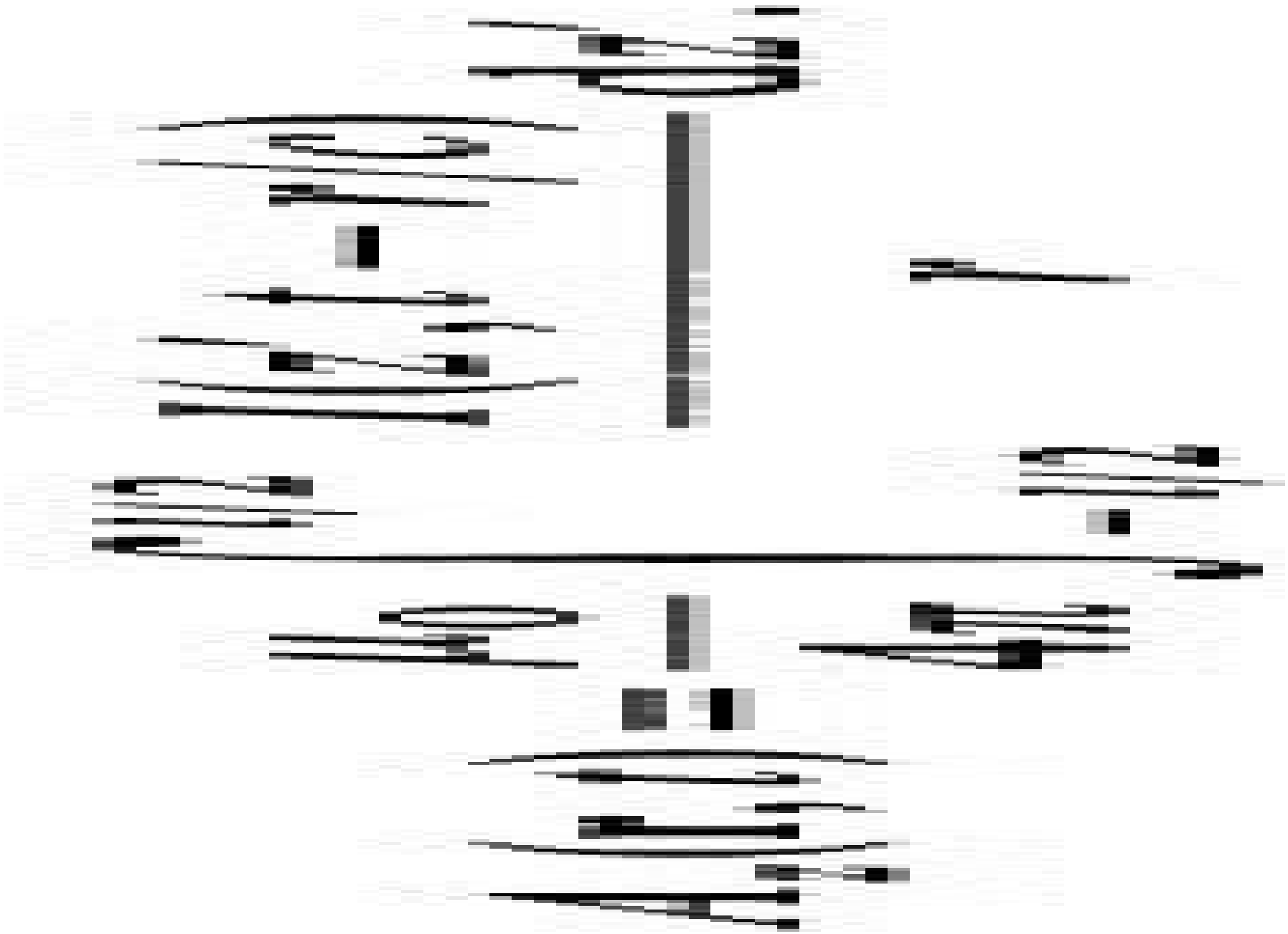
Let us consider the fields at points like the point P, far away from the dipole so that r » l. Under these circumstances the denominator in the integrand in Equation (13.15) can be put equal to r. The times $(t - |r - ezz'/c)$ at which the current density in the wire is to be determined can all be put equal to the same time $(t - r/c)$ if the biggest error in time Δt introduced by this approximation is very small compared with the period T of the oscillations. The biggest error in time is given by



which can be expanded to first order in l/r to give

(13.16)

where θ is the angle shown in Figure 13.2. The length l of the dipole is very much smaller than the wavelength of the radiation, i.e. $l \ll 2\pi c/\omega$. The period T is equal to $2\pi/\omega$, hence $l/c \ll T$, which from Equation (13.16), implies that the time Δt is indeed very small compared with the period T. Hence for points far away from the dipole Equation (13.15) can be written



Since the current is the same at all points along the short wire, this integral can easily be evaluated to give

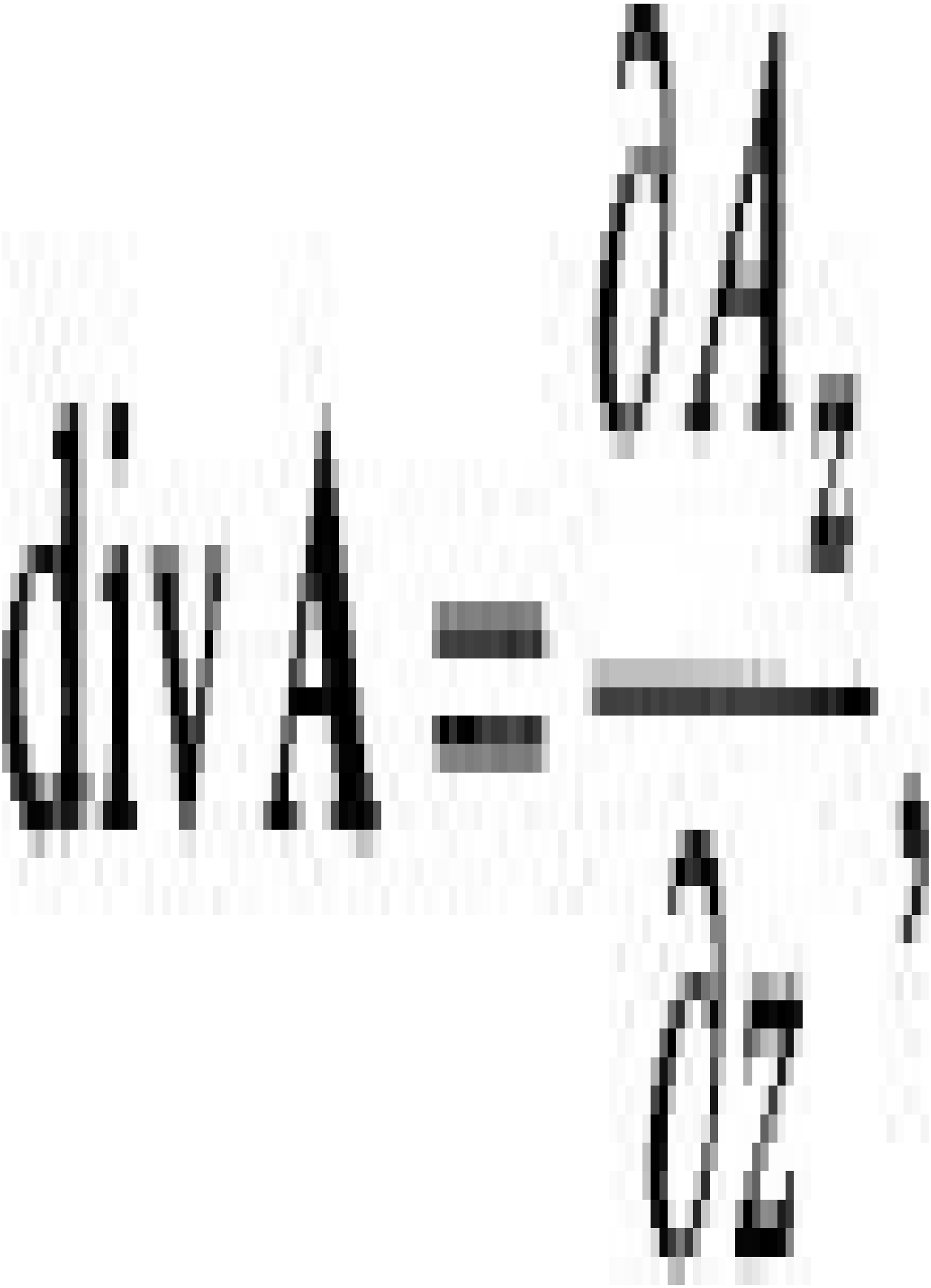
(13.17)

Since Ax and Ay are both zero, we now know the vector potential everywhere.

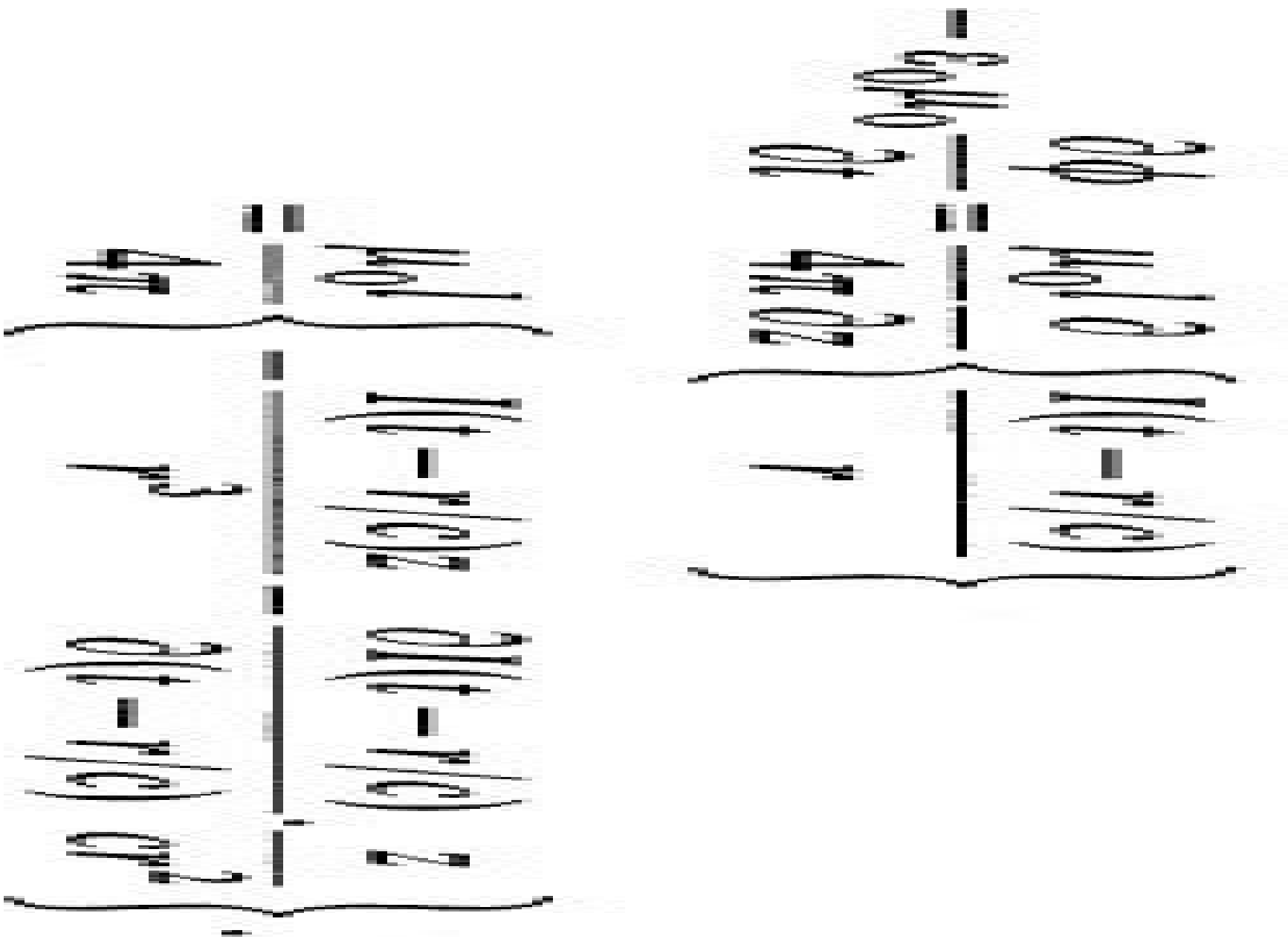
The potential ϕ can most easily be obtained from the Lorentz condition

(13.18)

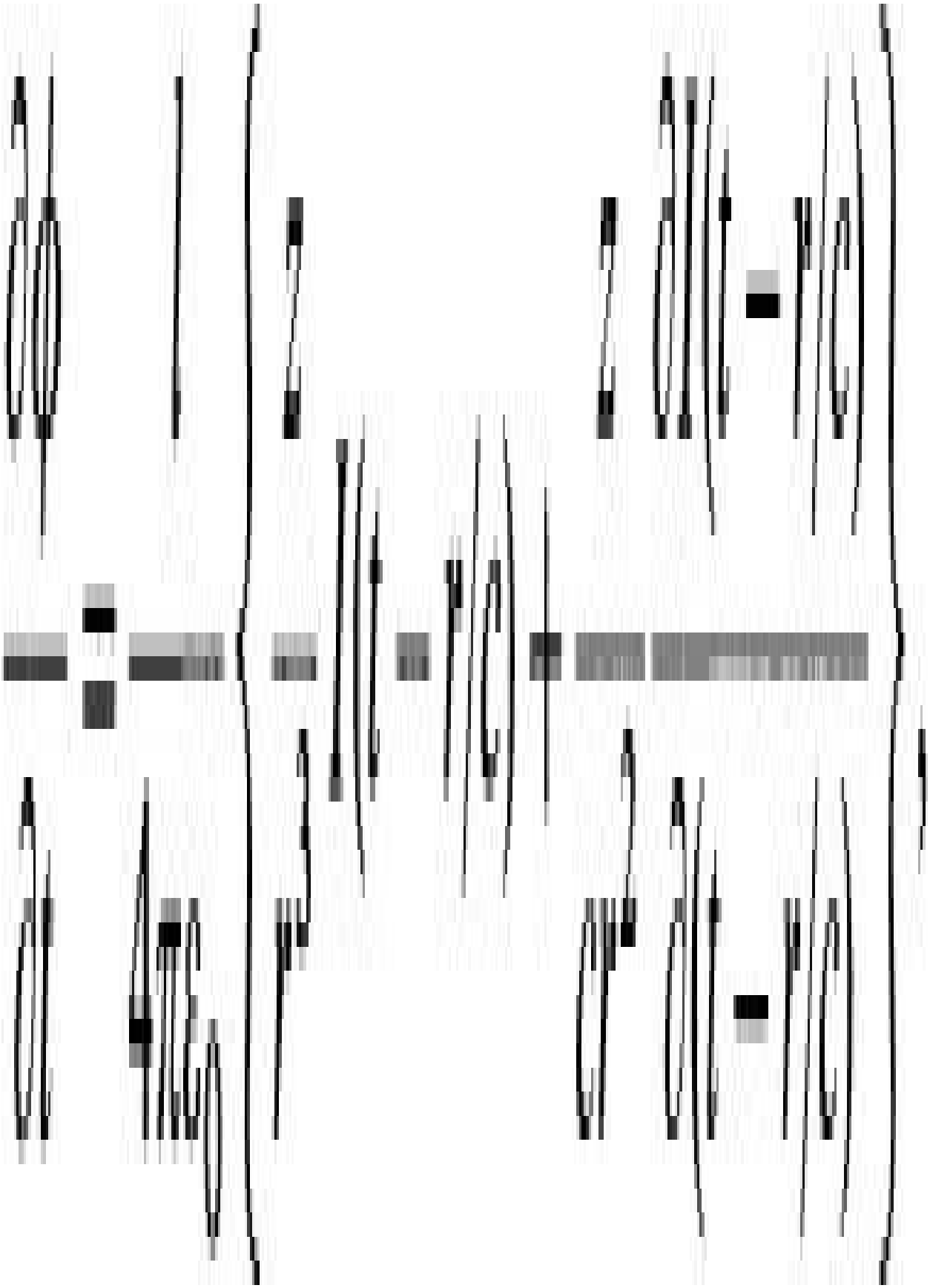
Since only the z-component of the vector A is non-zero



and by combining Equations (13.17) and (13.8) we obtain

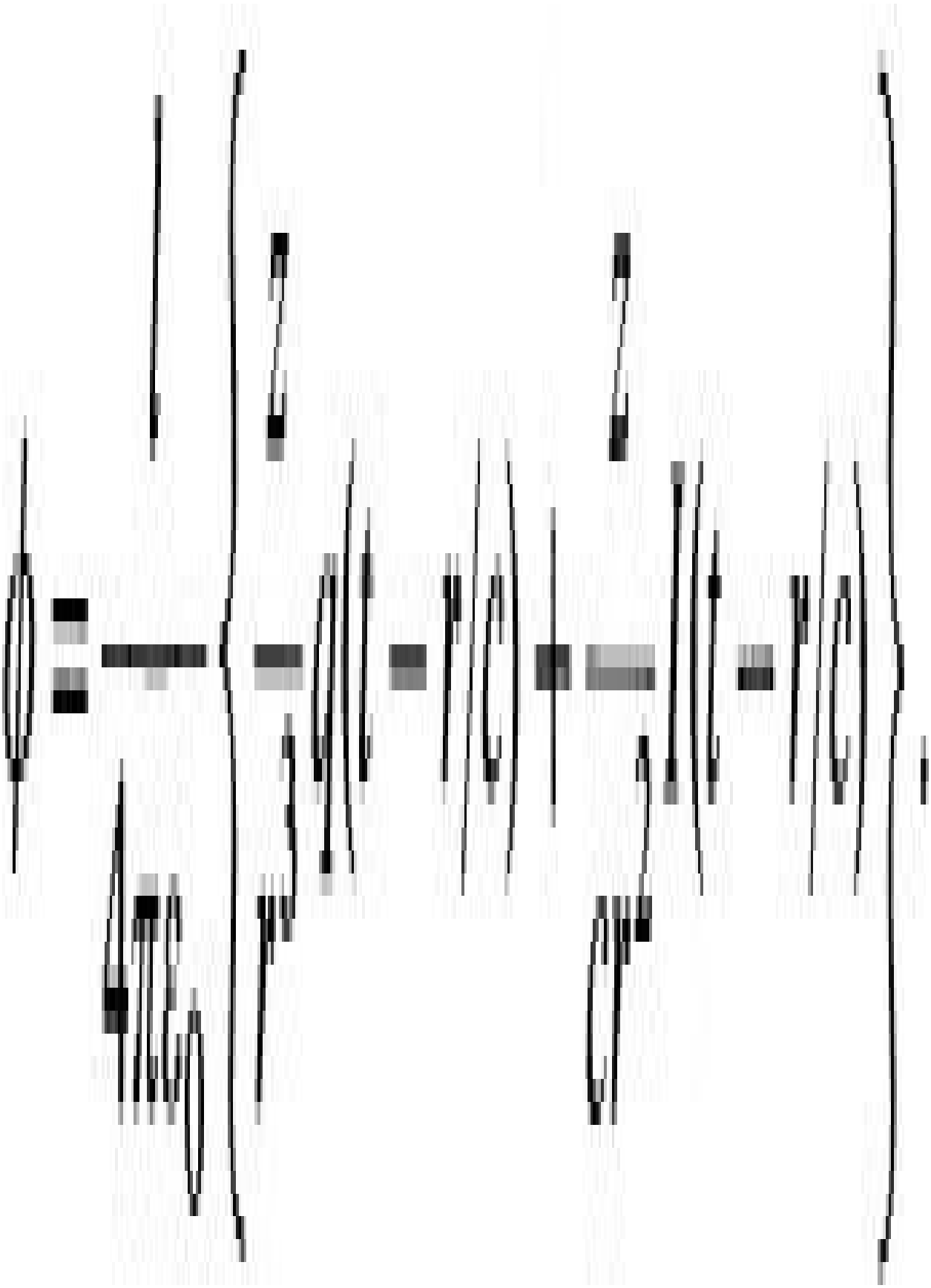


This reduces to the equation



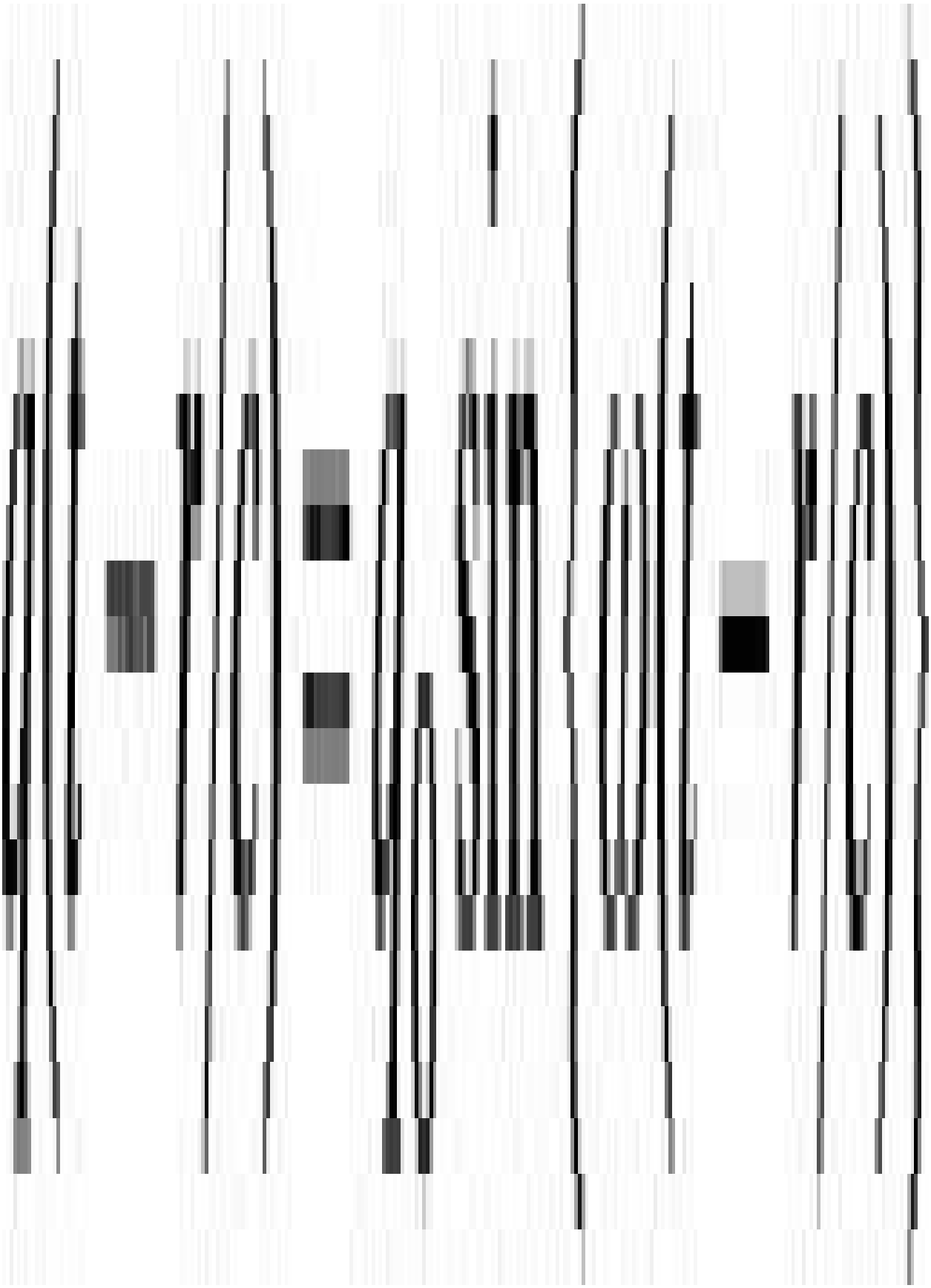
and on integration we obtain

(13.19)

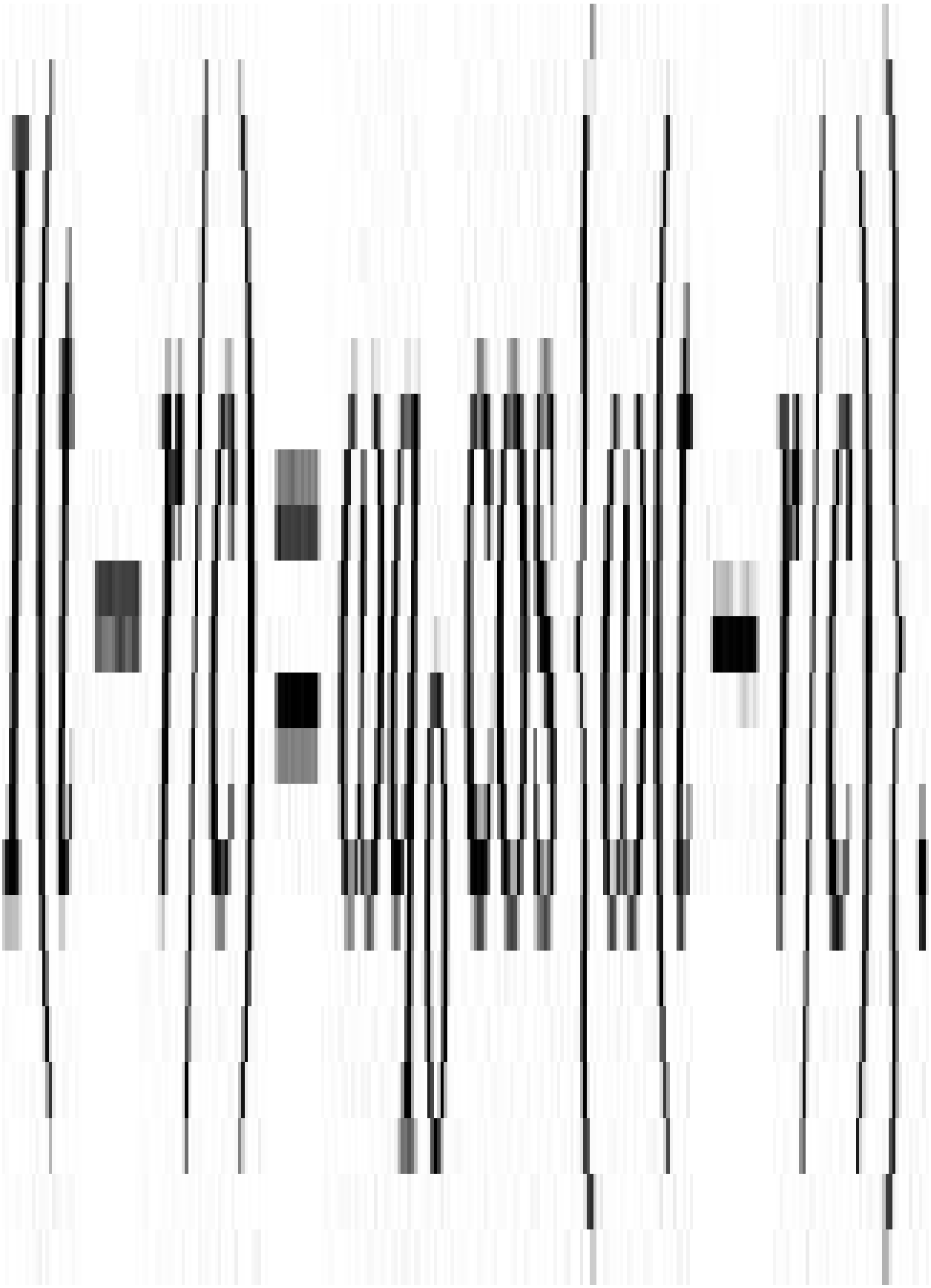


The forms of the potentials given by Equations (13.17) and (13.19) show that when the fields are calculated from them with the aid of Equations (13.1) and (13.2) there is a field component which falls off with distance proportionally to 1/r, plus other components which decrease more rapidly. The first component is the radiation field, the other terms are important for the fields near to the dipole. (Close to the dipole, at distances such that r l the potentials given above are not sufficiently accurate to determine the fields.)

Let us restrict attention to the radiation field far away from the dipole, where the other field components are negligibly small, and calculate the total power radiated and the manner in which the radiated energy is distributed in space. The fields E and B can be calculated from the potentials by writing



and



The first term on the right-hand side of Equation (13.19) can be ignored in calculating the radiation field, since it gives an electric field which decreases faster than 1/r. The obvious coordinate system to use in solving the problem is a spherical polar coordinate system. The expressions for the operators grad and curl in these coordinates are given in Appendix B. We will omit the mathematics of the derivation and quote the results. They are

(13.20)

$$E_r = 0$$

$$E_\phi = 0$$

$$E_\theta = -\frac{\omega l I_0 \sin\theta}{4\pi\varepsilon_0 c^2} \cdot \frac{\sin\{\omega(t - r/c)\}}{r},$$

(13.21)

$$B_r = 0$$

$$B_\theta = 0$$

$$B_\phi = -\frac{\mu_0 \omega l I_0 \sin\theta}{4\pi c} \cdot \frac{\sin\{\omega(t - r/c)\}}{r},$$

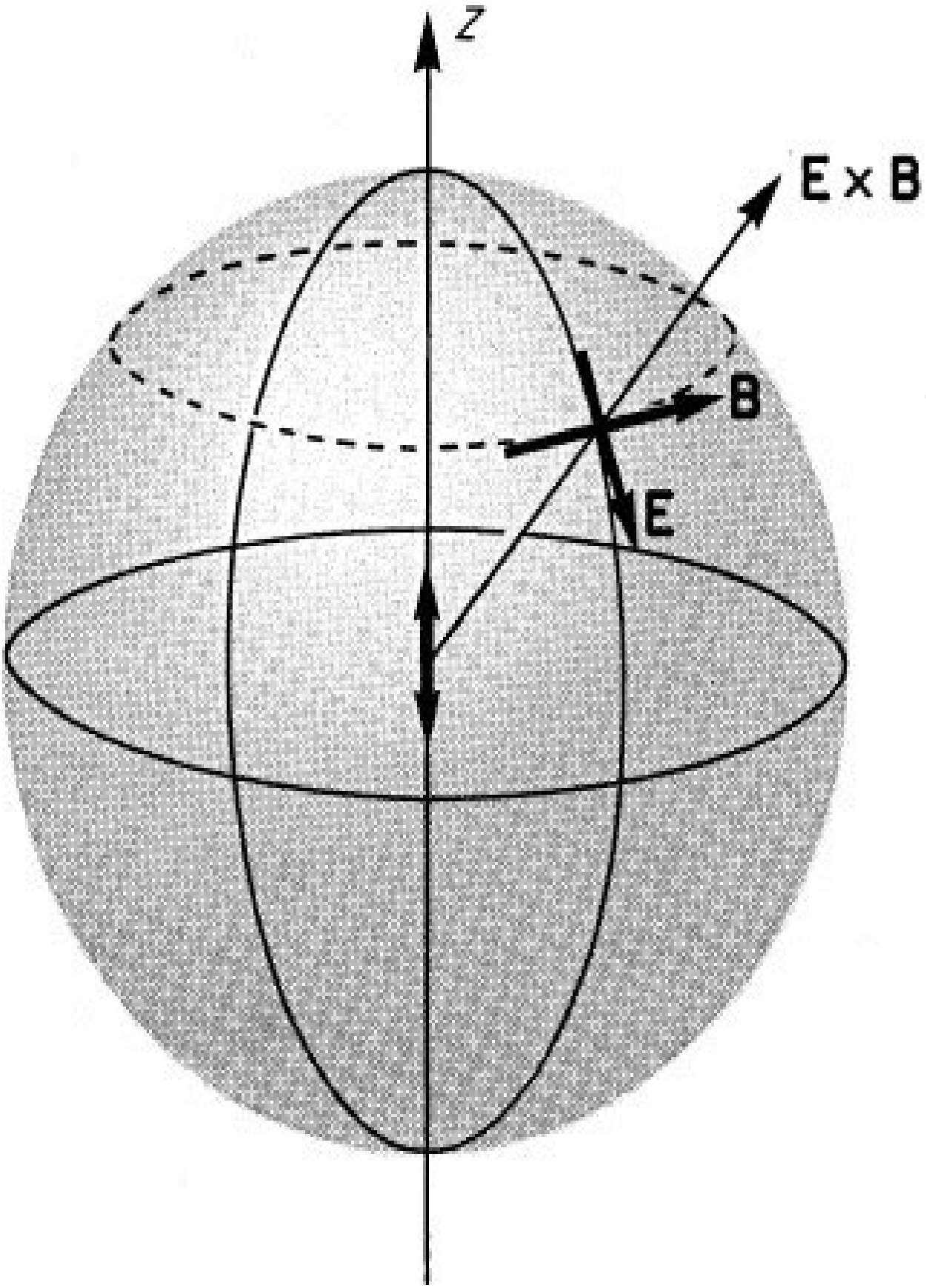
The relative directions of the fields are shown in Figure 13.3. They are mutually perpendicular and the energy flow at all points is in the direction of the radius vector. The fields are outgoing spherical waves and can be written (in a format similar to that used in Chapter 11 for plane waves) as the real parts of the expressions

$$E_\phi = \frac{j\omega l I_0 \sin\theta}{4\pi\varepsilon_0 c^2} \cdot \frac{\exp\{j(\omega t - kr)\}}{r}$$

$$B_\phi = \frac{j\omega l I_0 \sin\theta}{4\pi\varepsilon_0 c^2} \cdot \frac{\exp\{j(\omega t - kr)\}}{cr}.$$

The fields are symmetric about the z-axis and their magnitudes are independent of the azimuthal angle ϕ. This must be so since the dipole looks the same from all points with the same values of the coordinates r and θ. The fields are zero at all points on the axis of the dipole, and rise to maximum values on a disc that passes through the middle of the dipole and cuts it perpendicularly, i.e. at $\theta = \pi/2$. The spatial distribution of the fields can be shown on a polar diagram, such as Figure 13.4, which gives the relative values of the fields at different positions on the surface of a sphere centred on the dipole. For example the ratio of the electric fields at two points on the sphere with polar angles $\theta 1$ and $\theta 2$ is given by the ratio of the lengths of the lines joining the origin O to the points P1 and P2 on the polar diagram.

Figure 13.3. The relative directions of the radiation fields of a Hertzian dipole. The fields are tangential to the surface of a sphere centred on the dipole. The electric field oscillates in the direction of the unit vector $e\theta$, and the magnetic field in the direction of the unit vector $e\phi$.

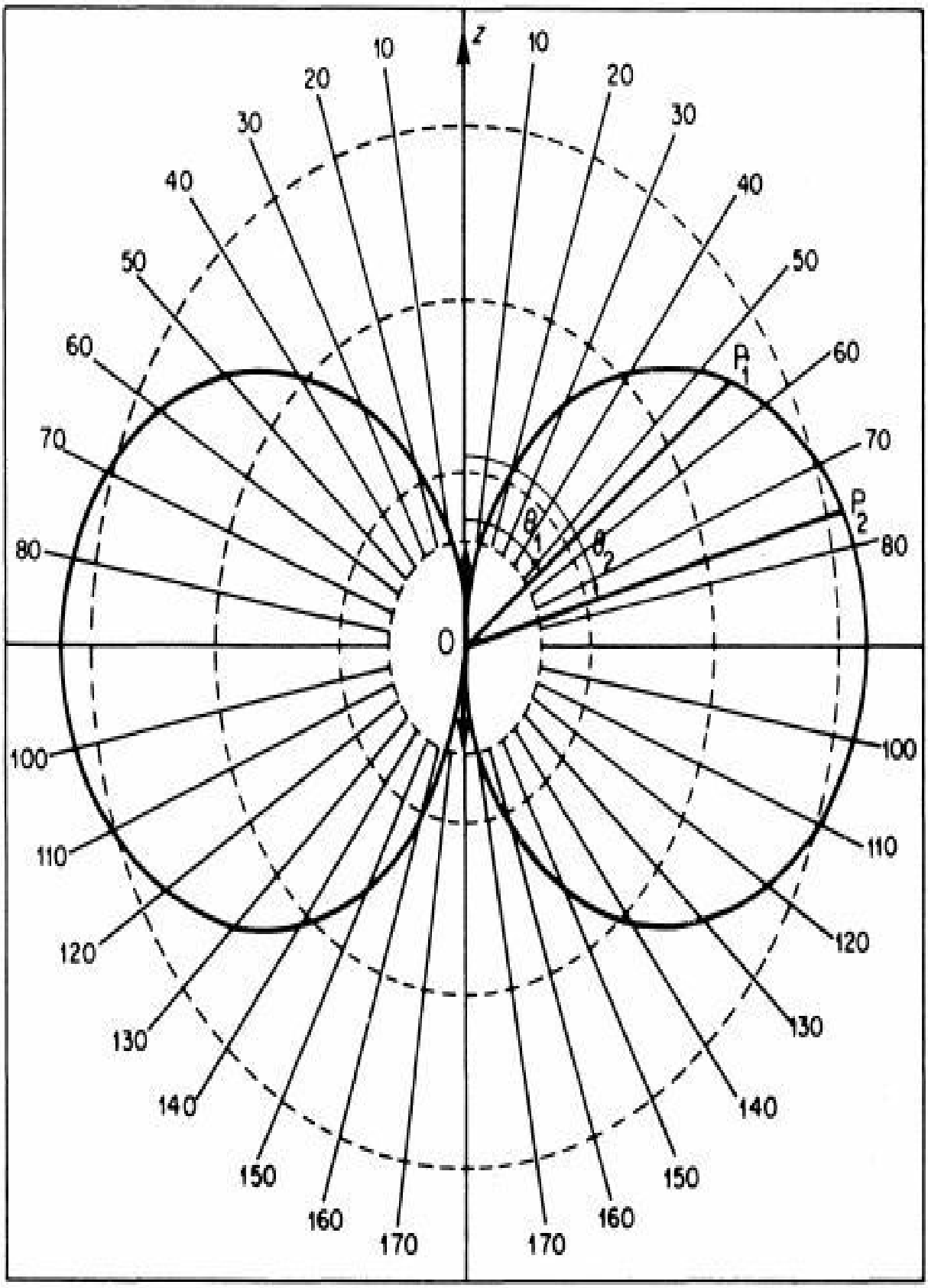

Z
E x B
B
E

The instantaneous total power W crossing the surface S of any sphere of radius r in the radiation zone is given by integrating the Poynting vector N over the surface.

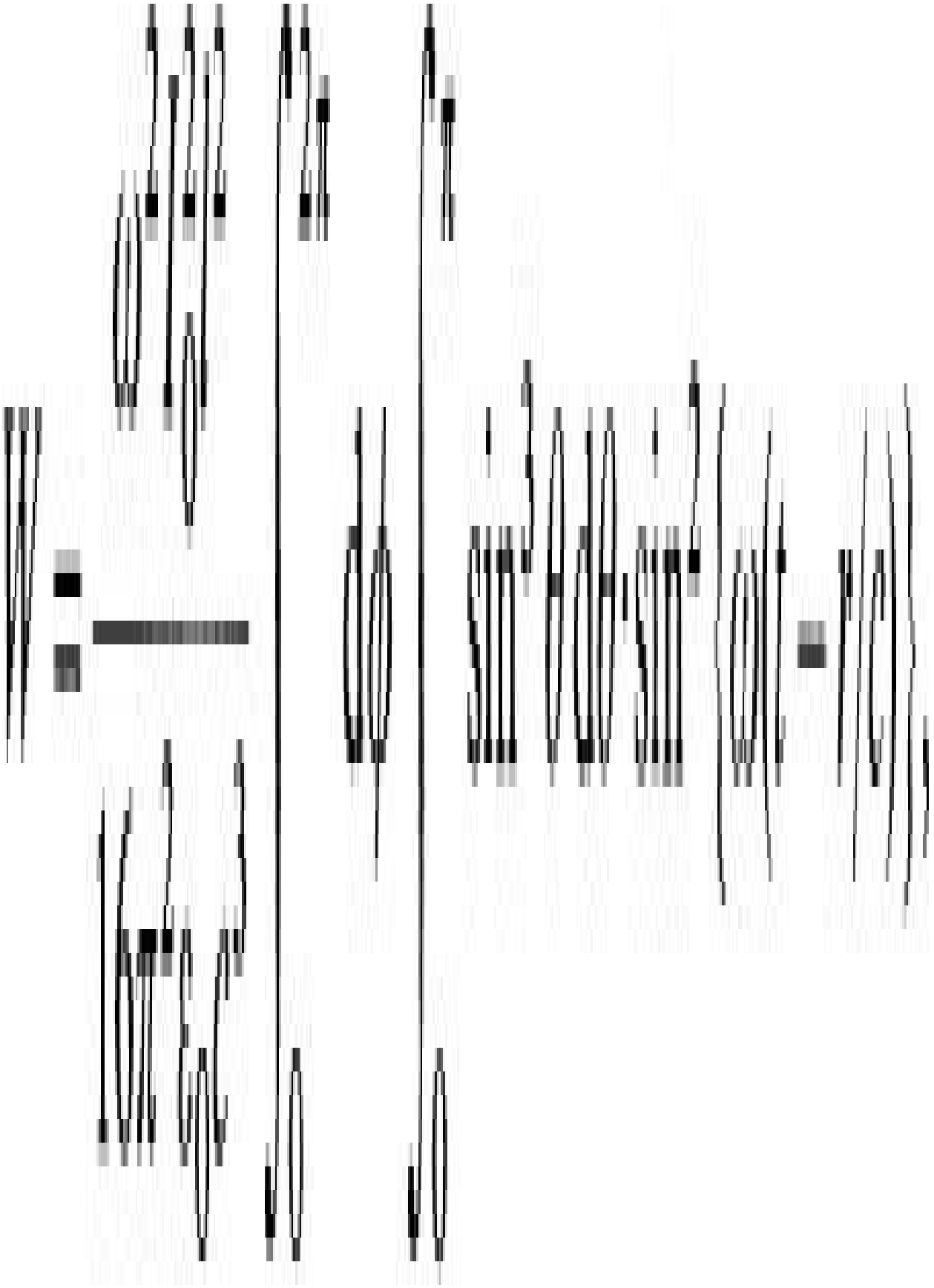
$$W = \oint_S N \cdot dS$$

$$= \frac{1}{\mu_0} \int_0^{2\pi} \int_0^{\pi} B_\phi E_\theta r^2 \sin\theta \, d\theta \, d\phi.$$

Figure 13.4. The polar diagram for a Hertzian dipole. The solid curved lines give the relative strengths of the radiation field at different points on the surface of a sphere centred on the dipole. The dipole lies along the z-axis and the field is independent of the azimuthal angle ϕ, depending only on the polar angles $\theta 1$ and $\theta 2$ of the points on the sphere.


z
10
20
30
40
50
60
70
80
100
110
120
130
140
150
160
170
0
P1
P2
θ1
θ2

Hence



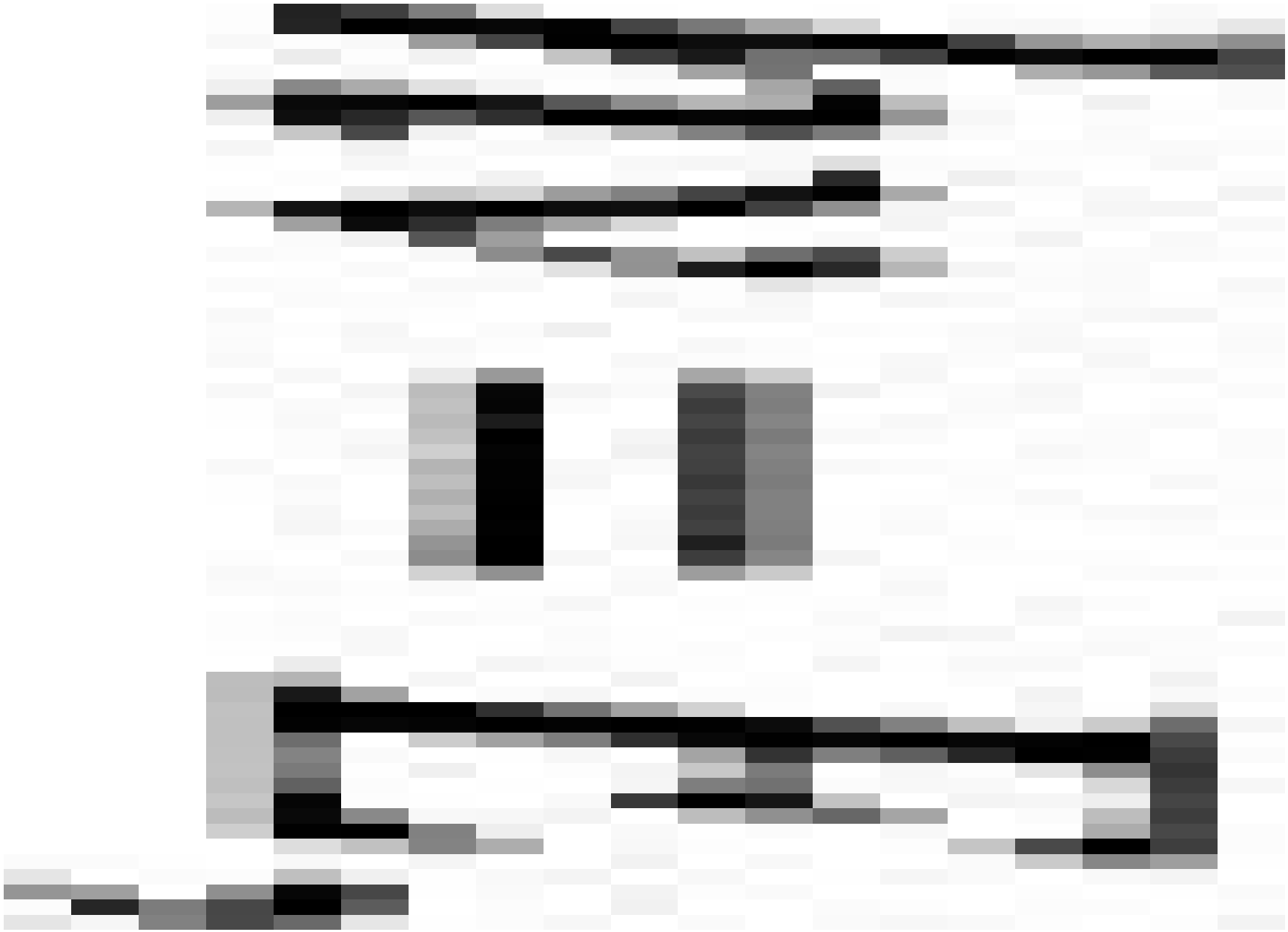
from Equations (13.20) and (13.21) The average power radiated is given by averaging this expression over a cycle of the oscillations:

(13.22)

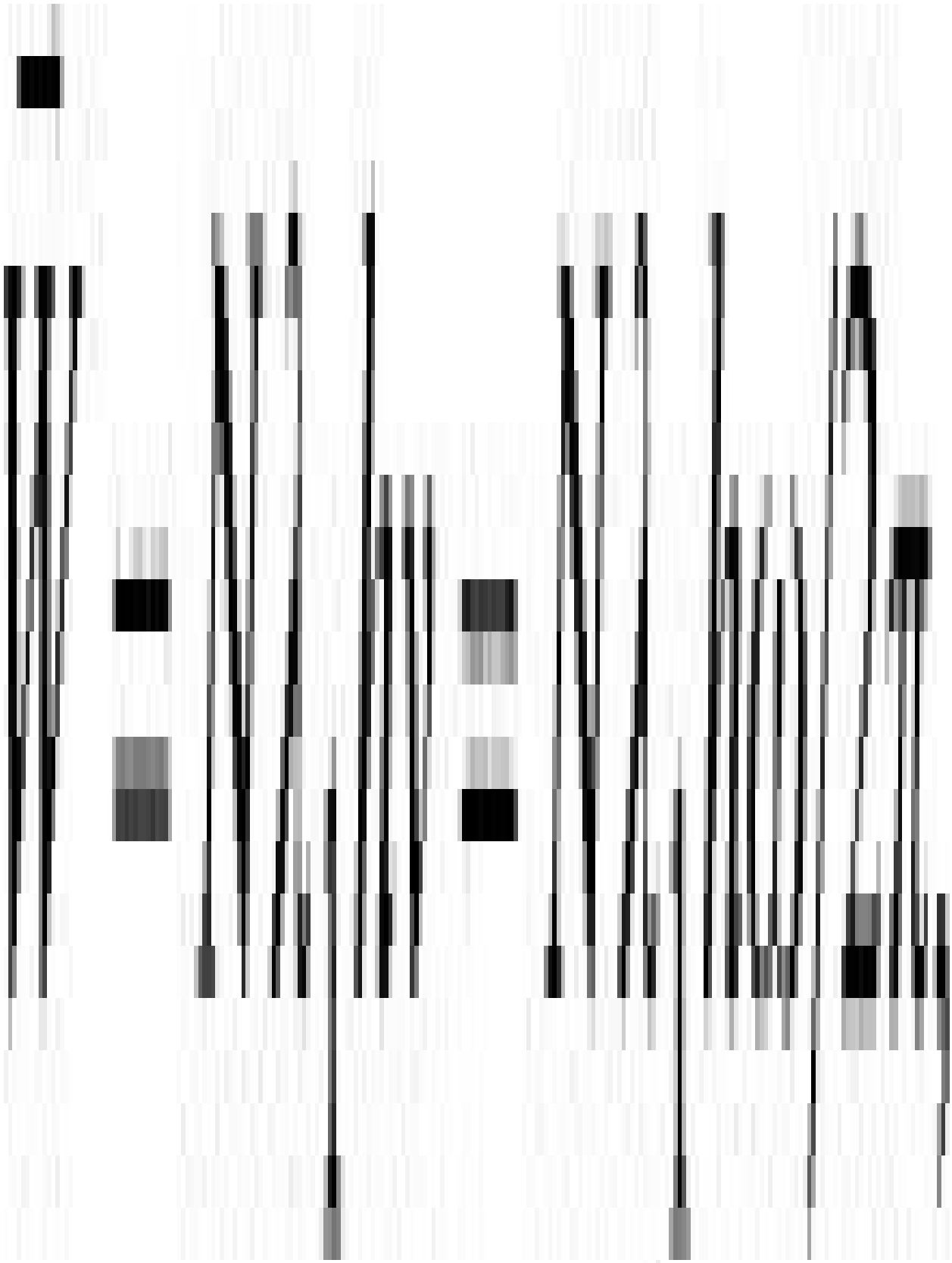
This can be rewritten in terms of the amplitude p0 of the oscillating dipole moment, using Equations (13.13) and (13.14):

(13.23)

As mentioned earlier in this section, the results given above for the Hertzian dipole are useful in the quantum theory of the radiation from microscopic particles. An atom in an excited state readjusts its charge distribution at some instant and may emit a packet of energy in the form of electromagnetic radiation as the atom reverts to its stable ground state. This packet of energy takes away the original excitation energy E of the atom, and is called a photon. The excited state of the atom has a fixed probability per unit time λ1 that it will decay by emission of a photon of energy E and leave the atom afterwards in its ground state. If there are N excited atoms the number of decays per second to the ground state is thus Nλ1. Hence the energy emitted per second in the decays to the ground state is equal to Nλ1E. The frequency v of the electromagnetic wave associated with photons of energy E is given by the formula



where h is Planck’s constant. The power radiated in the ground state decays of the N atoms is thus



where ω is the angular frequency of the radiation. We may deduce an expression for the decay constant λ1, by making an analogy between the change in the atomic charge distribution and a classical Hertzian dipole. The dipole moment p0 in Equation (13.23) is replaced for the atomic radiators by a factor M which can be calculated using quantum mechanics, and depends on the structure of the two atomic states involved in the transition. Equation (13.23) then predicts that the power radiated by the N atoms making transitions to the ground state is given by

$$W = \frac{NM^2\omega^4}{12\pi\varepsilon_0 c^3}$$

Equating this to the earlier expression for gives

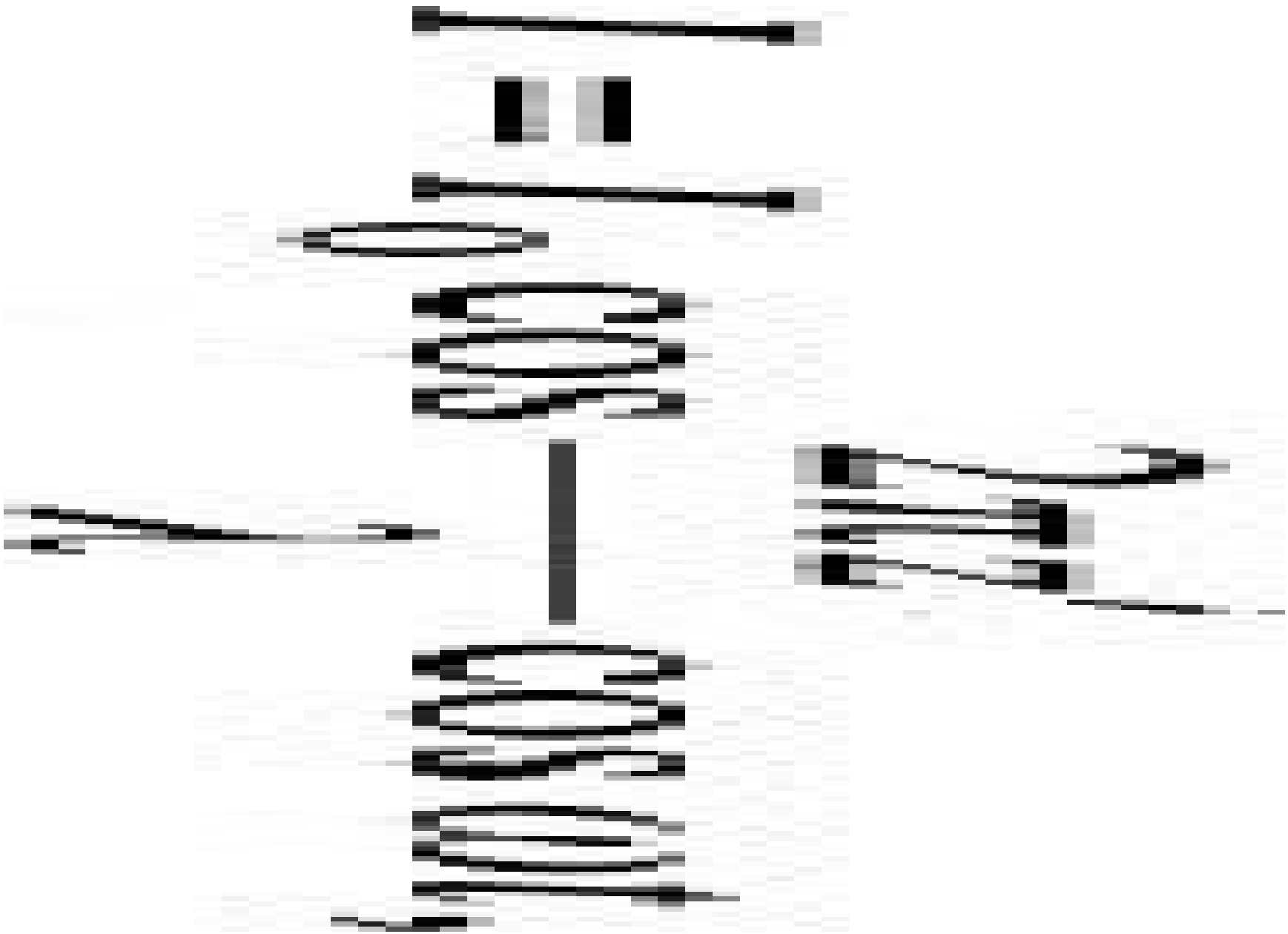
$$\lambda_1 = \frac{M^2\omega^3}{6\pi\varepsilon_0 c^3}$$

This expression shows that the decay probability $\lambda 1$ is proportional to the cube of the energy of the photon released in the decay. This result is also obtained in a more rigorous quantum mechanical treatment of atomic electric dipole transitions.

13.3 ANTENNAS

The calculation of the last section for the small dipole is not applicable to aerial systems used for transmitting radio waves. Usually these aerials have lengths which are not short compared to the wavelength λ of the radiation they transmit, and the variation of the amplitude of the oscillating current along the aerial (antenna) has to be taken into account.

A simple antenna normally consists of a straight wire or metal tube an integral number of half-wavelengths long. The antenna has capacitance and inductance, and can carry current. It behaves like a length of open-ended transmission line: standing waves are set up, with the currents satisfying the condition that they be zero at each end of the wire. The current is usually fed into the aerial via a coaxial cable transmission line. Figure 13.5 shows schematically the arrangement of a transmitter which has a half-wave dipole antenna. The current I in this antenna at a point distance z′ from the centre is given to a good approximation by the expression



where ω is the angular frequency of the oscillations.

The radiation field from such an antenna can be determined by splitting it into infinitesimally small elements, each of which can be considered to be a Hertzian dipole. The contribution $dE\theta$ to the electric field at a far off point P from the small dipole of length dz' at a point Q on the aerial with coordinate z' is given by Equation (13.20) as

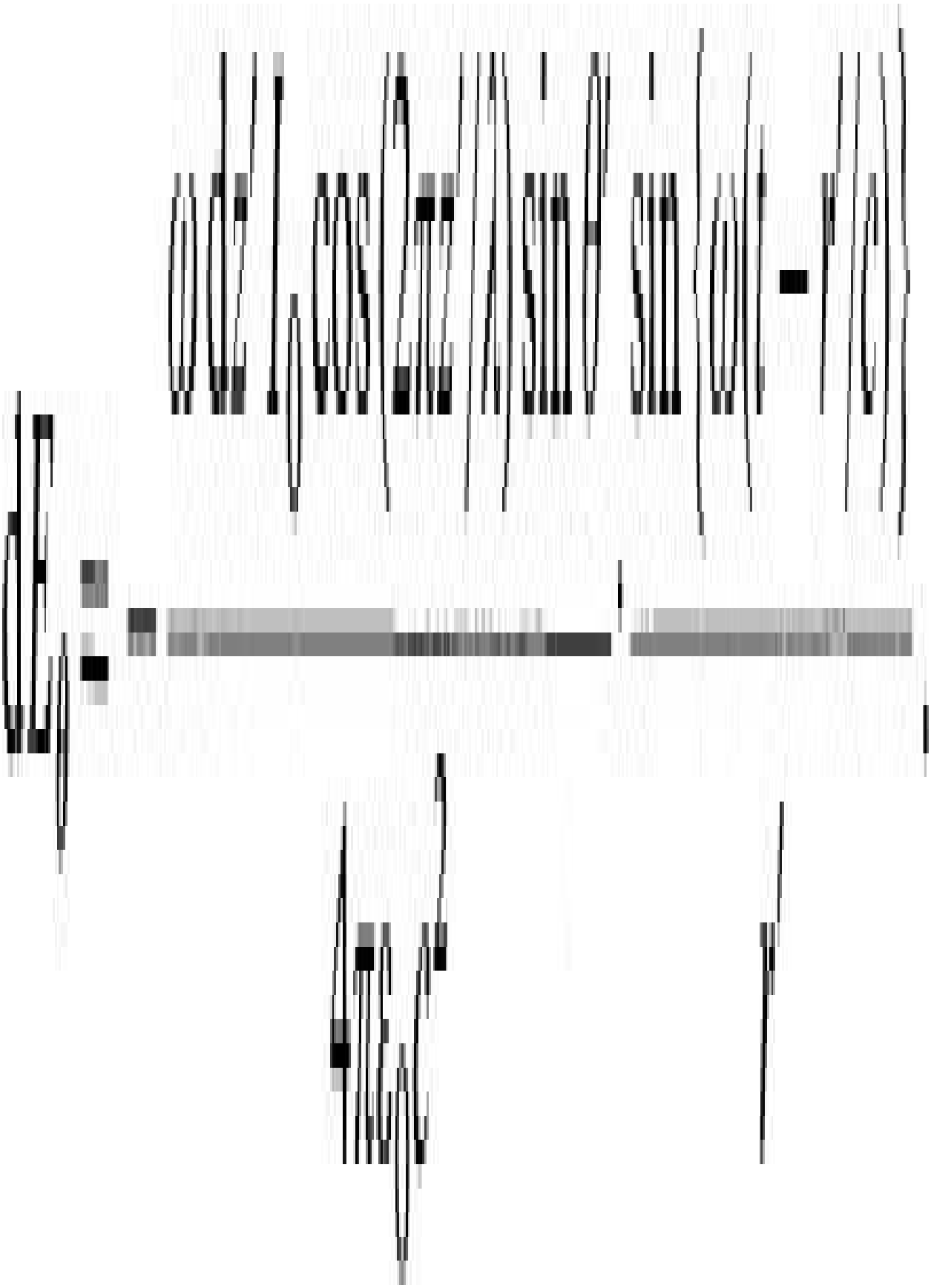
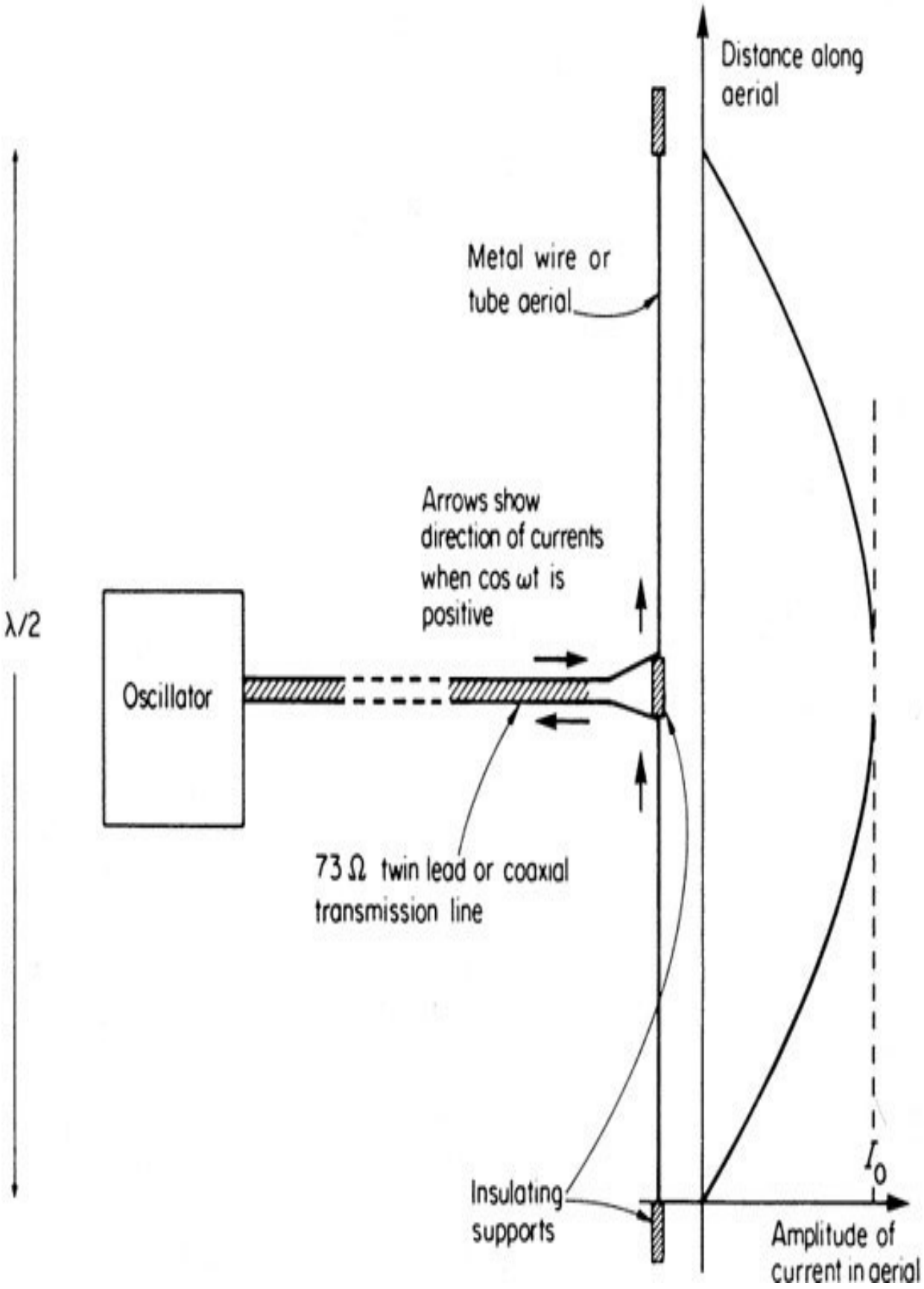
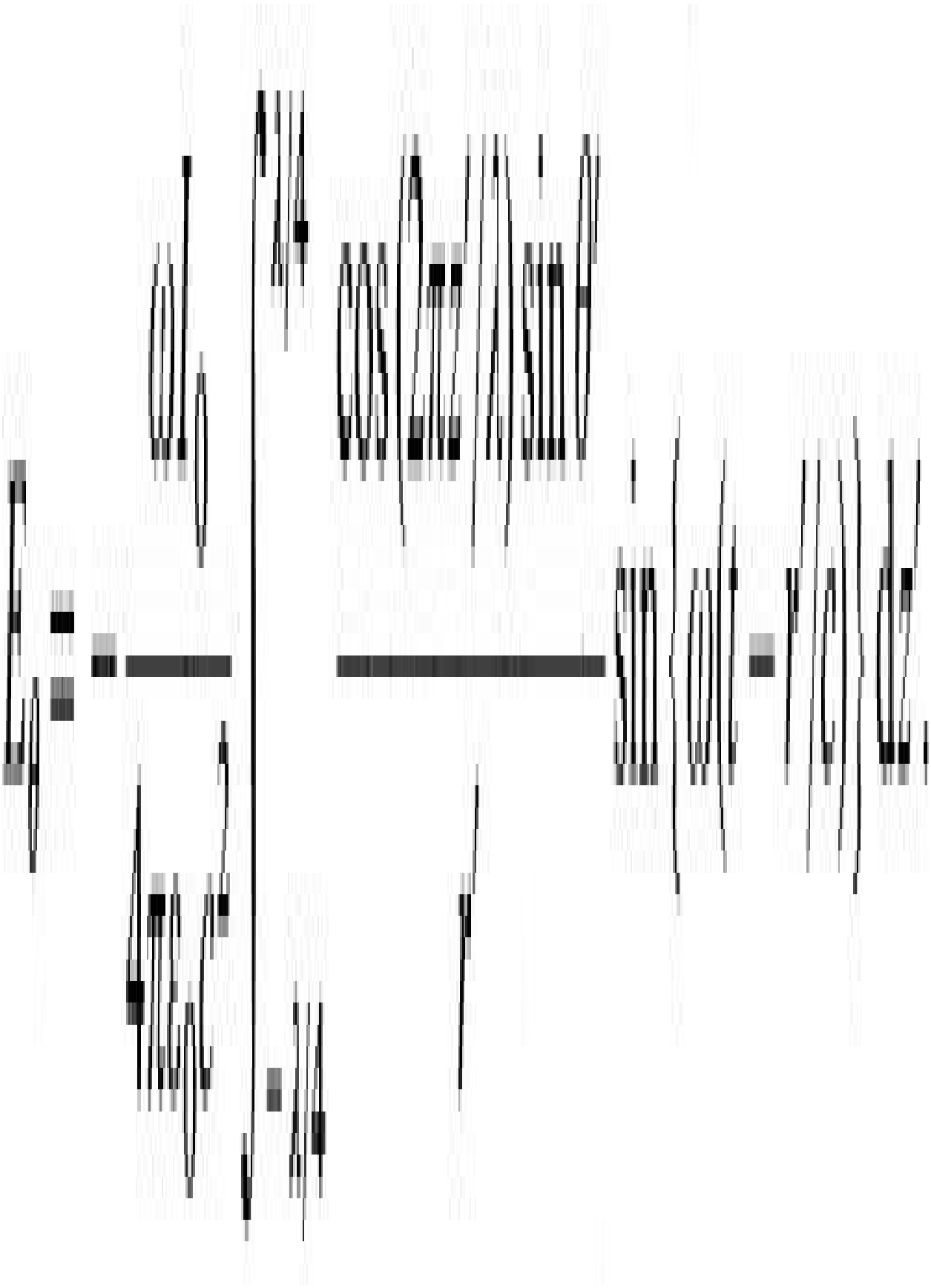


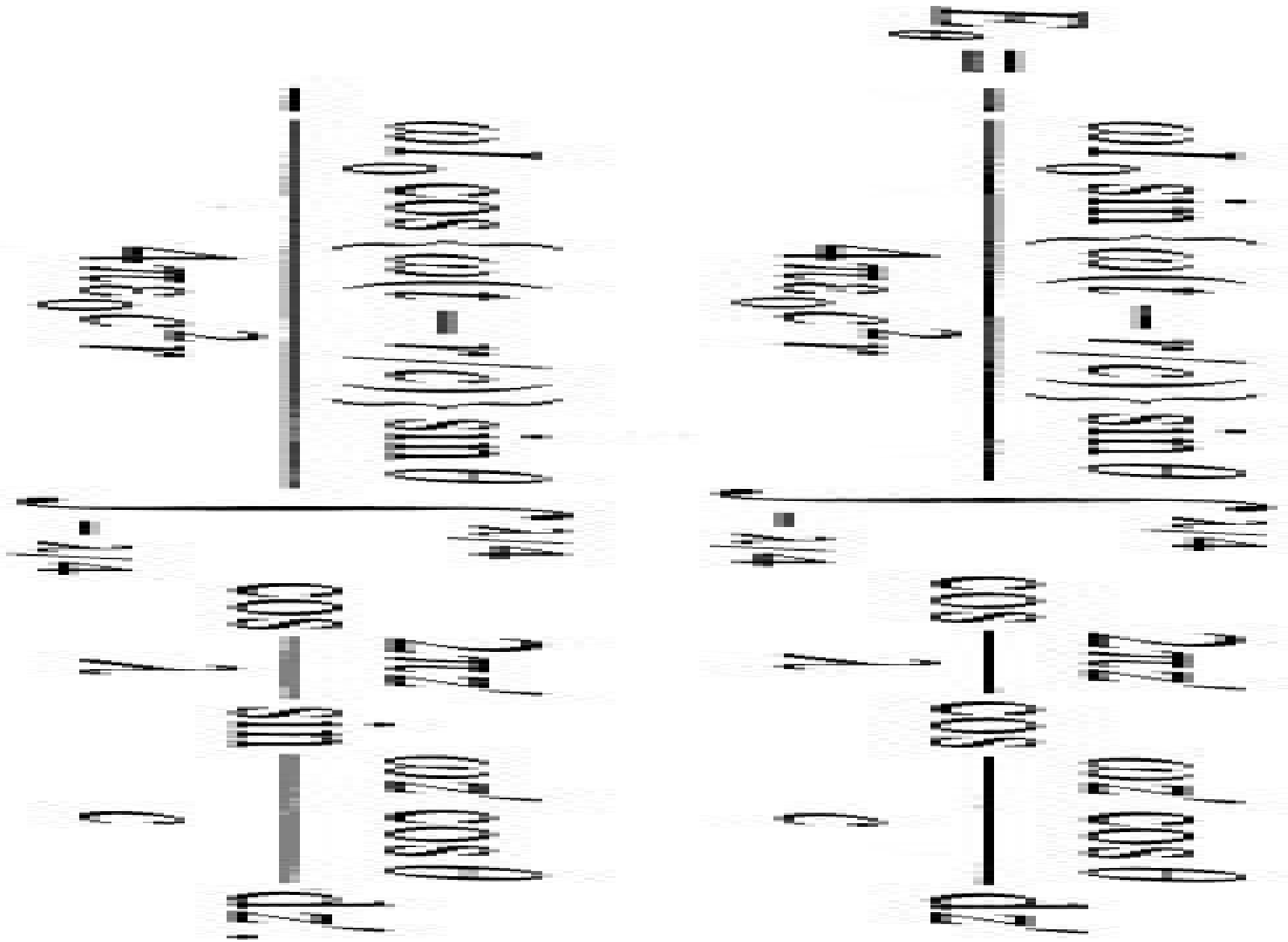
Figure 13.5. A schematic arrangement of a transmission line coupling an oscillator to a half-wave dipole aerial. The current amplitude varies approximately sinusoidally along the aerial.


Distance along aerial
Metal wire or tube aerial
Arrows show direction of currents when cos ωt is positive
λ/2
Oscillator
73 Ω twin lead or coaxial transmission line
Insulating supports
I_0
Amplitude of current in aerial

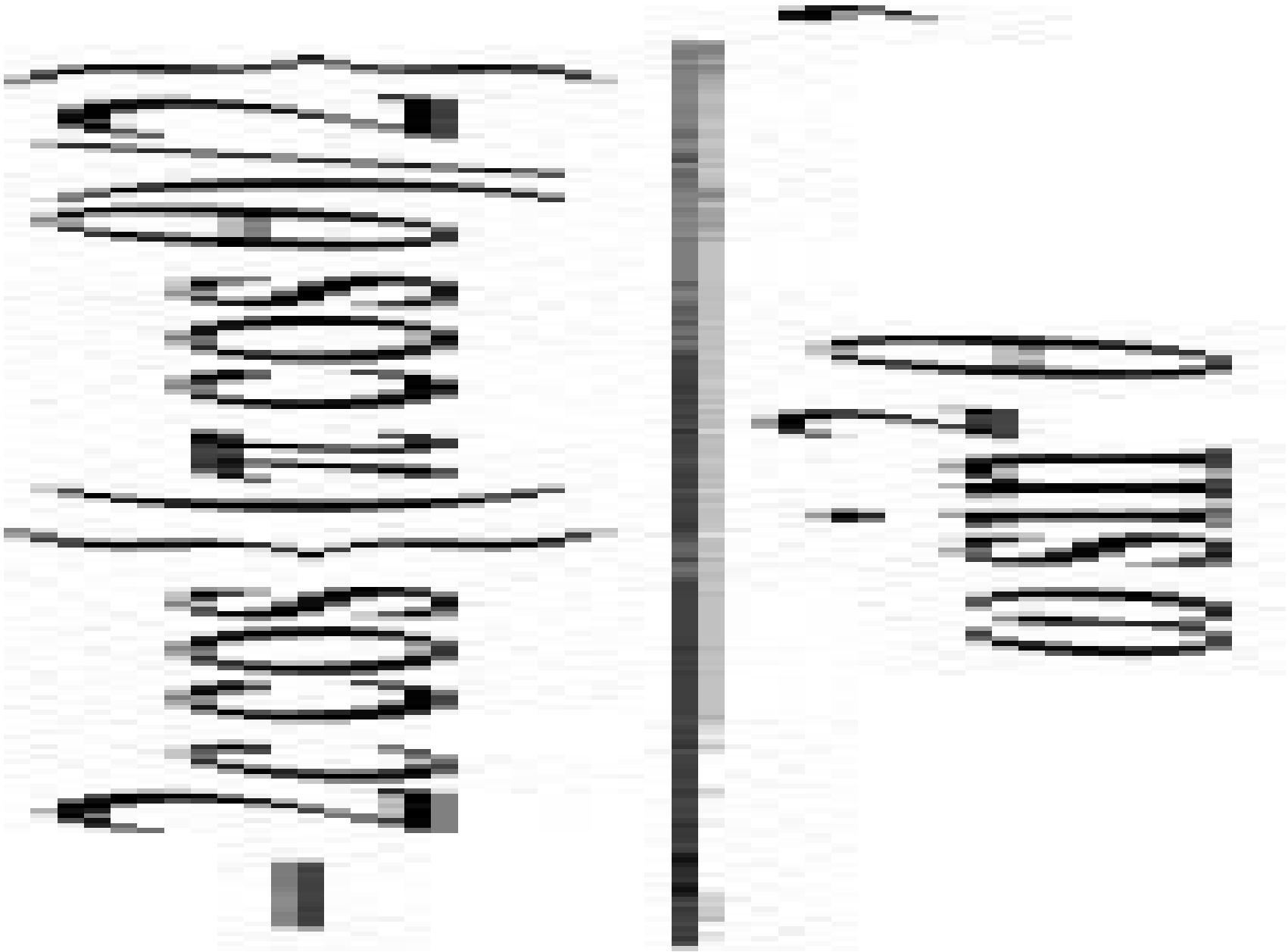
In this equation r' is the distance from the point Q to the point P and θ' is the angle the line joining Q to P makes with the z-axis. The z-axis is taken to be along the aerial and we assume $r' \gg \lambda$. The total electric field $E\theta$ at P is obtained by integrating over the length of the aerial, and we have



To first order in z'/r, the distance $r' = r - z' \cos\theta$ where r is the distance from the centre of the dipole to the point P. Since $r \gg \lambda$ the term $\sin\theta'/r'$ in the integrand can be replaced by the constant $\sin\theta/r$. The phase factor $\omega(t - r'/c)$ has to be retained however, since the time z'/c is now comparable with the period of the oscillations. Hence

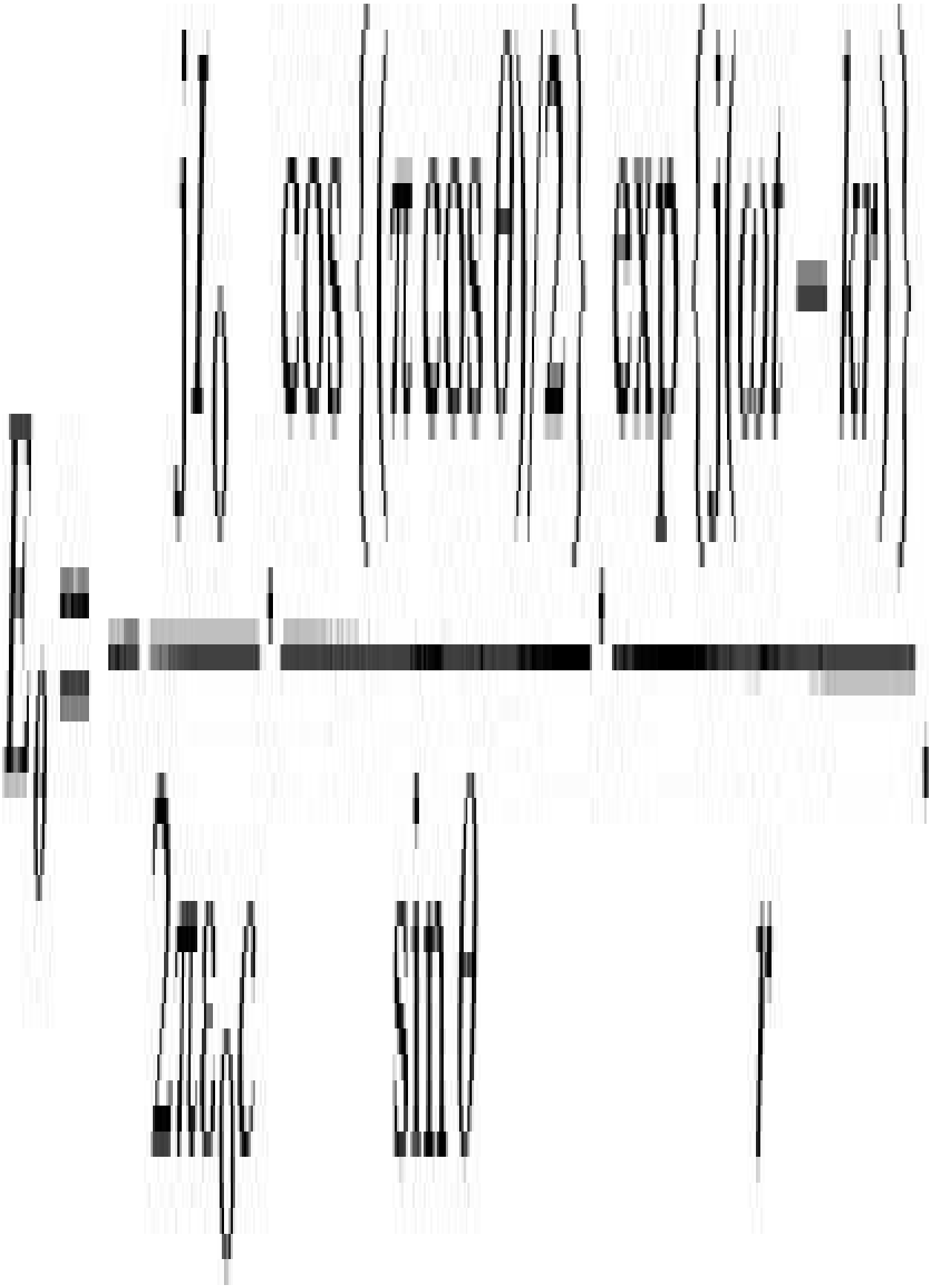


The second integral vanishes. After some straight forward but lengthy algebra the first integral can be shown to be equal to



and the field at P is thus given by the real part of the expression

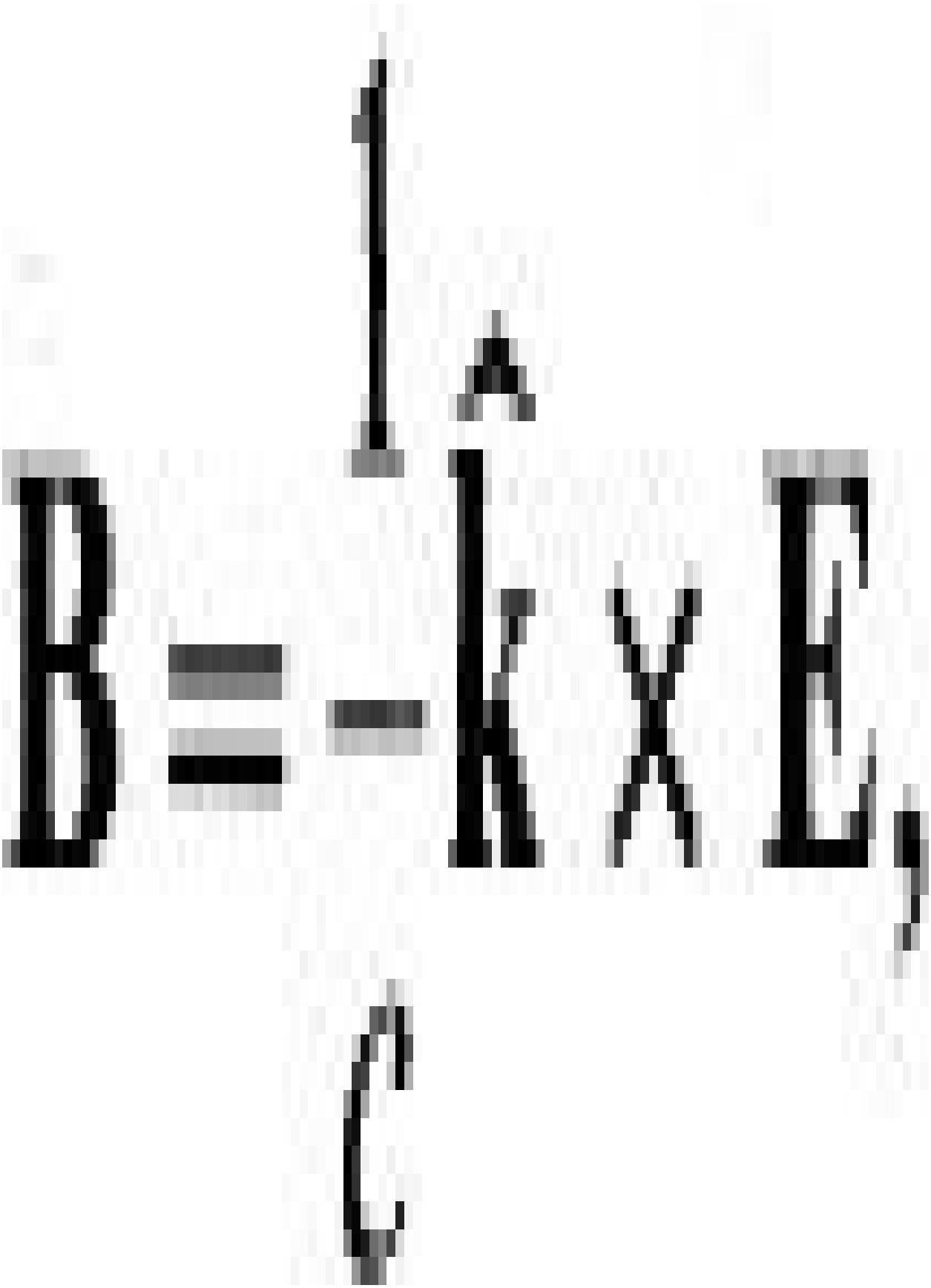
(13.24)



The magnetic field $B\phi$ is given by the equation

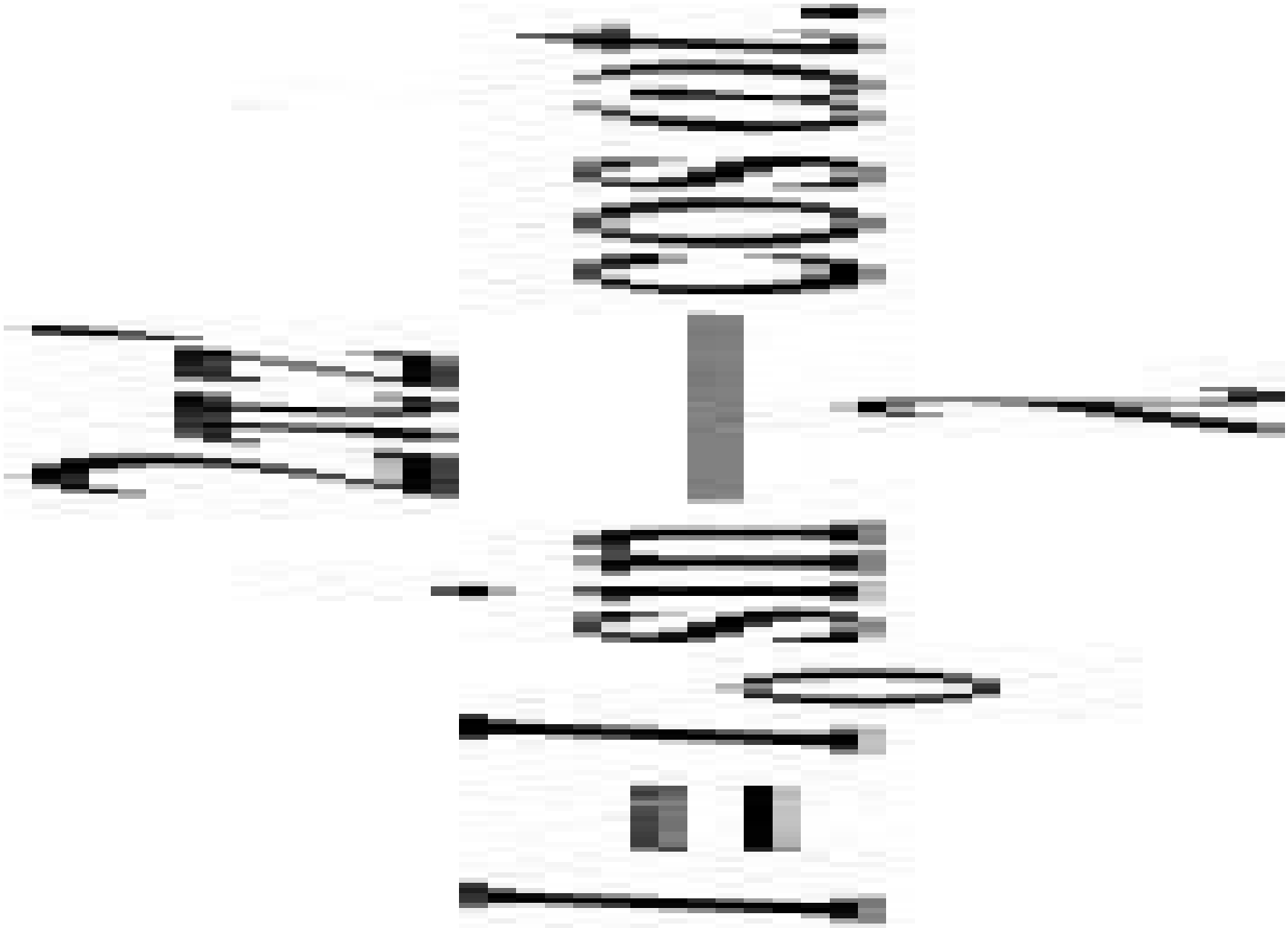
(13.25)

as can be seen from Equations (13.20) and (13.21). This is another illustration, like the plane wave example, of the general rule that the relationship between the magnetic and electric fields in an electromagnetic radiation field in free space is



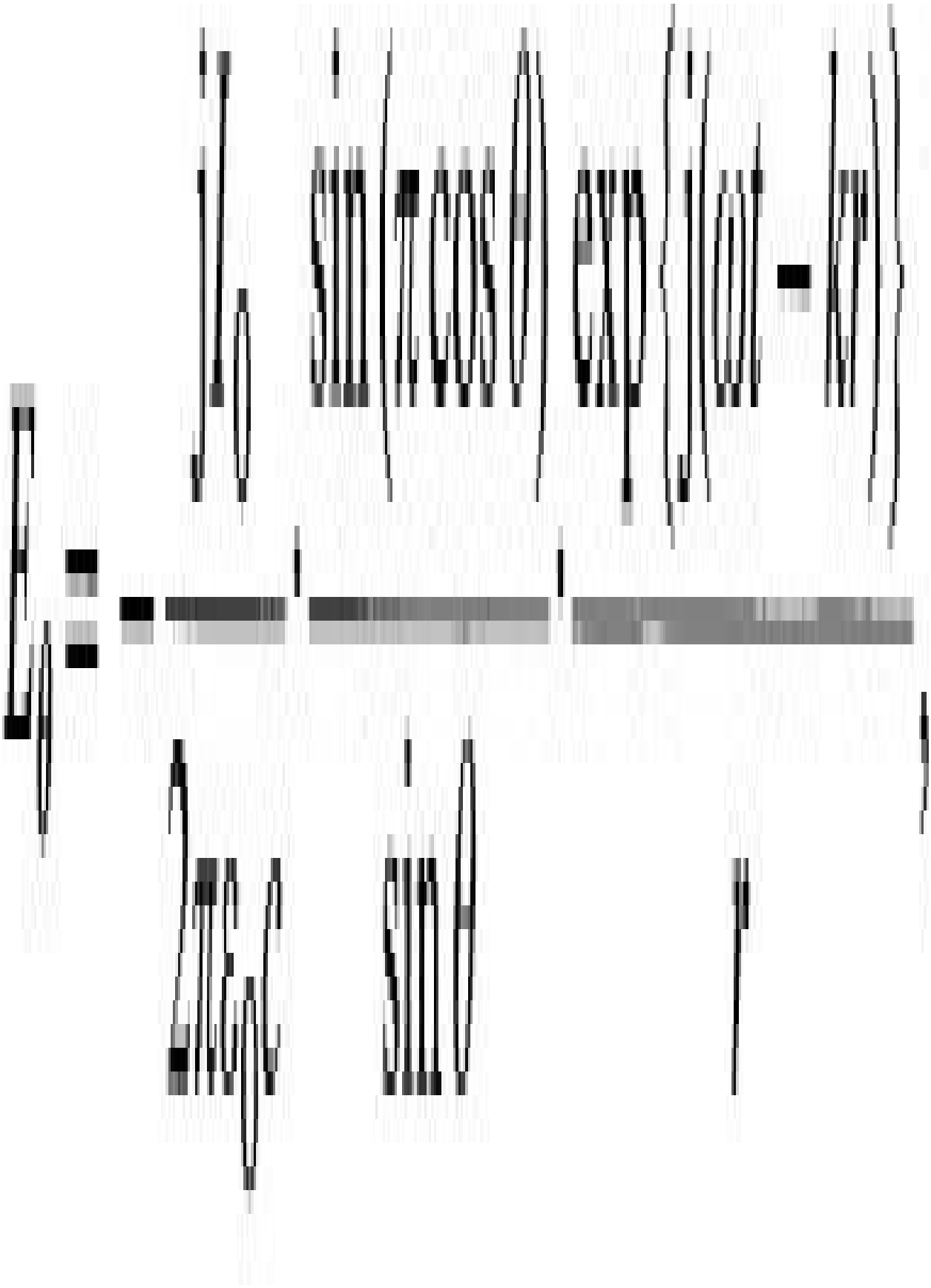
where is a unit vector in the direction of propagation of the wave.

For a full-wave aerial, one wavelength long, the current at a point with z-coordinate z' is



The z-axis is taken to be along the aerial with the origin at its centre as before. A similar procedure to that followed with the half-wave aerial gives the following expressions for the fields:

(13.26)



and

(13.27)

The polar diagrams for the fields from half-wave and full-wave aerials are shown in Figures 13.6 and 13.7.

The average total power radiated by a half-wave aerial can be obtained by integrating the Poynting vector over the surface of a complete sphere.

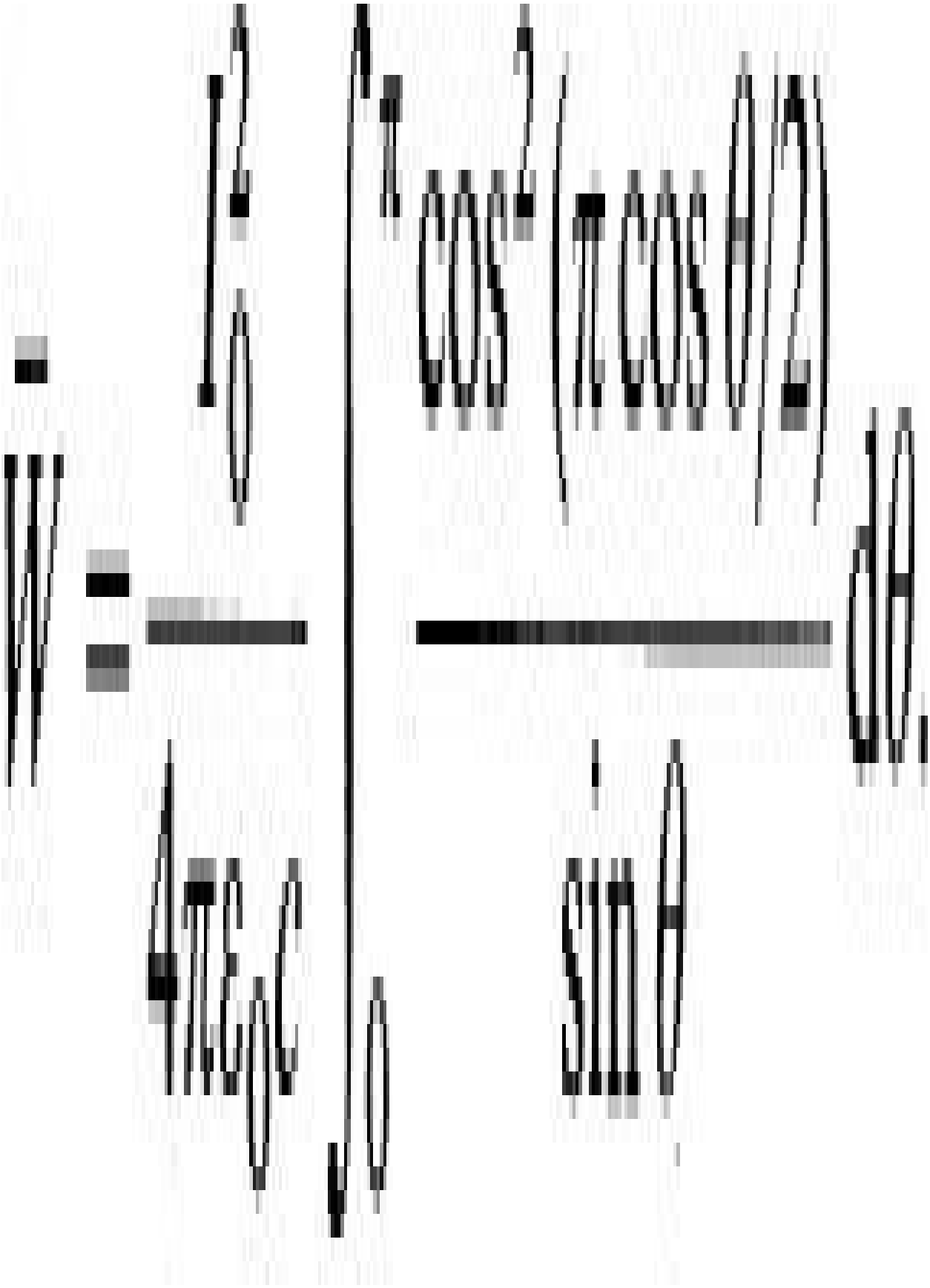
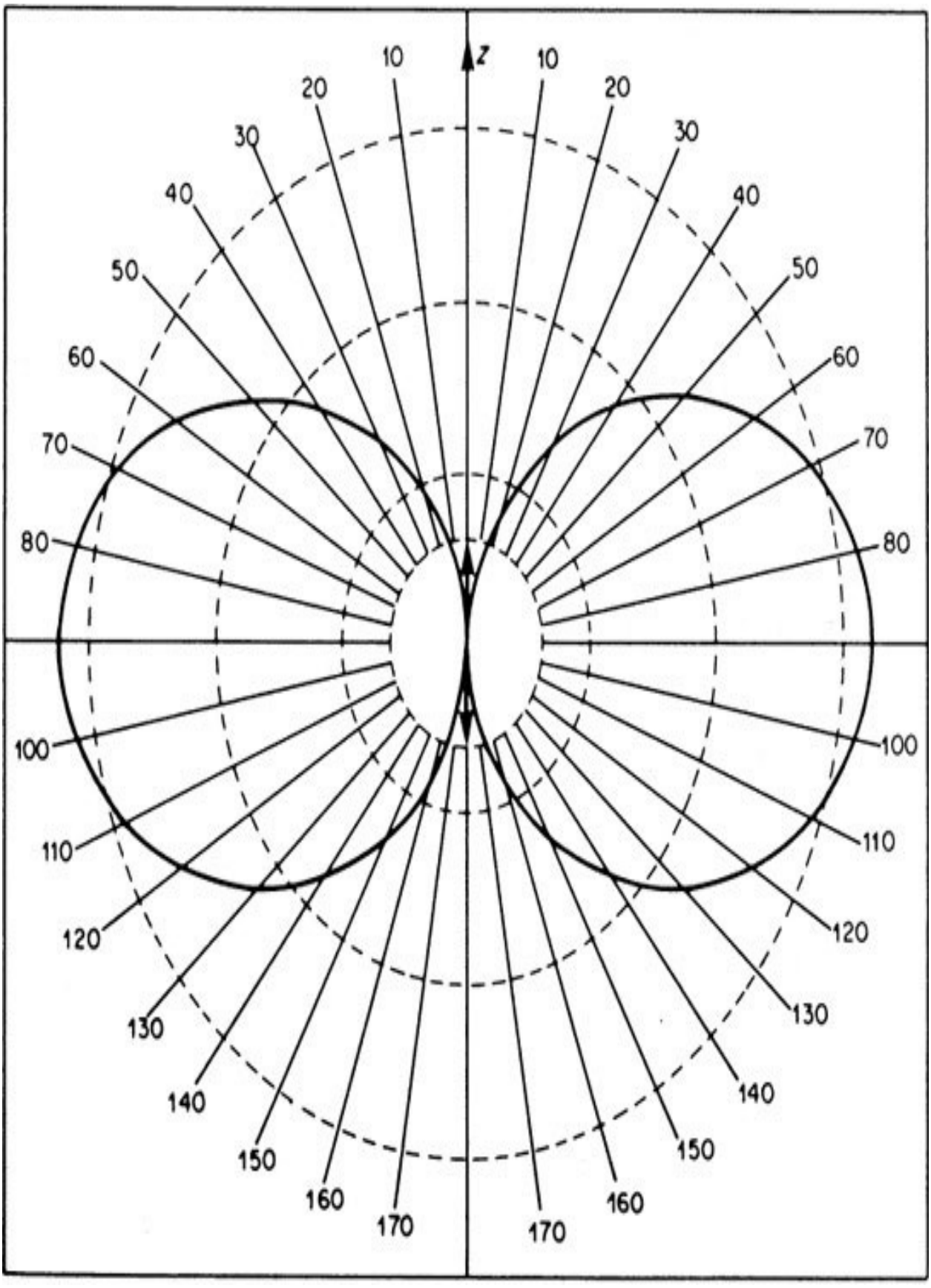


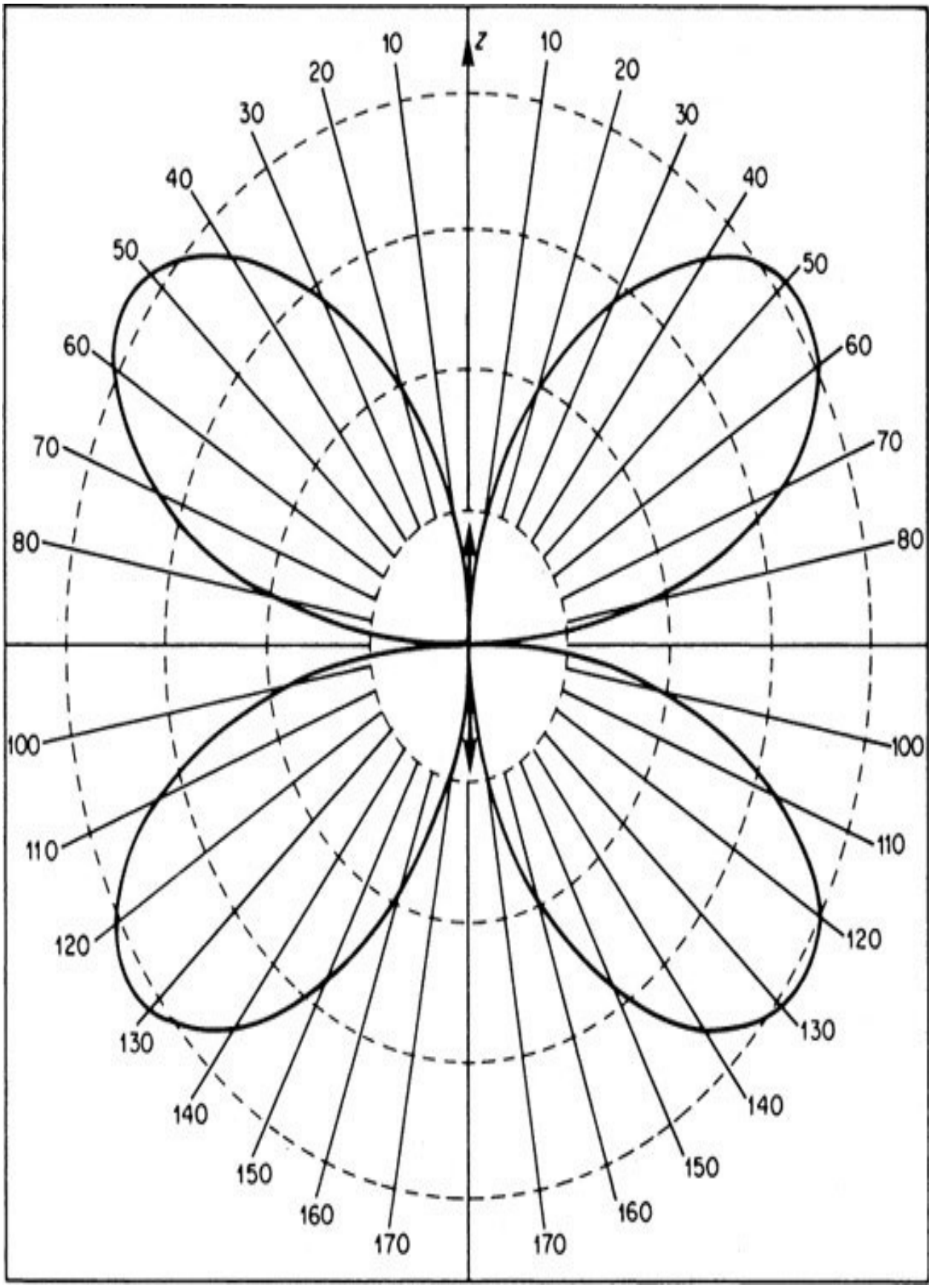
Figure 13.6. The polar diagram for a half-wave dipole aerial.


z
10
20
30
40
50
60
70
80
100
110
120
130
140
150
160
170

The integral can be evaluated numerically to give

In order to transmit 5 kW from a half-wave aerial the amplitude of the current oscillating in it at its centre must be about 12 A. If an oscillating current of amplitude I0 were passing through a resistor of resistance 73 Ω the average power dissipated would be watts. The radiation resistance of the half-wave antenna is thus 73 Ω, and is a convenient quantity used to represent the aerial in an equivalent circuit.

Figure 13.7. The polar diagram for a full-wave dipole aerial. Note the second zero at θ = 90°. This occurs because the full-wave aerial is similar to two half-wave aerials placed end to end carrying out of phase currents.


z
10
20
30
40
50
60
70
80
100
110
120
130
140
150
160
170

The field patterns and other properties we have worked out above apply for isolated aerials, those which are many wavelengths clear of the ground or other bodies which will reflect electromagnetic waves. These patterns are almost impossible to obtain in practice, and at frequencies below about 3 MHz the height of the antenna above ground is a major factor in determining its radiation pattern. The optimum antenna system for a given requirement is usually determined empirically, and the final radiation pattern measured experimentally. The matching of the antenna to the transmission line feeding it is also usually done empirically, using theory as a guideline. A half-wave dipole aerial looks like a real impedance of about 73 Ω to the transmission line, since the ohmic losses in the metal antenna are negligible compared to the energy radiated. The exact value of the equivalent impedance of the antenna again depends on several factors, among which are the height of the dipole above the ground, the point on the dipole to which the transmission line is joined, and the diameter of the dipole. A transmission line with characteristic impedance different from that of the aerial can be matched to the half-wave dipole in several ways, one of which is by using a length of transmission line of suitable characteristic impedance, as discussed in section 9.4.

An antenna system can be made more directional by using more than one dipole radiator. For example if two half-wave dipoles are placed parallel, side by side and separated by a short distance, the resultant field is the sum of the fields due to each separately. At some points the fields will be out of phase and will cancel each other. At other points the fields will interfere constructively. The aerial system will be directional, and the directional property can be varied by altering the relative phases of the currents fed to the two dipoles. For example consider two half-wave dipoles pointing vertically and separated in the horizontal planes by a distance equal to one half of a wavelength, as in Figure 13.8. If the same currents are fed to them in phase, the field at a particular point is given by the sum of two terms which will be the same except for a phase difference corresponding to the different path lengths from the dipoles to the point under consideration. Along the x-axis the fields everywhere cancel out, whilst they are always in phase along the y-axis. The relative fields in the x–y plane are shown on Figure 13.9. The use of several dipoles instead of just two would enhance the directional properties of the antenna even further. The radiation pattern produced

by a system consisting of a large number of dipoles fed with their currents all in phase is analogous to the diffraction pattern produced when a plane wave of monochromatic visible light is incident normally onto a diffraction grating*. Another way of obtaining a highly directional antenna is to situate a half-wave dipole at the focus of a parabolic reflector. This produces a narrowly diverging beam whose direction can be simply varied by movement of the system as a whole. The angular half-width of the beam θ is that appropriate to the diffraction pattern of a circular aperture* of diameter d equal to the diameter of the parabolic reflector. θ is thus roughly equal to (λ/d) where λ is the wavelength of the radiation.

Figure 13.8. A simple antenna consisting of two vertical half-wave dipoles separated by a distance equal to one half of a wavelength. The connections to the transmission line coupling the antenna to the oscillator are made so that the currents in both dipoles are in phase.

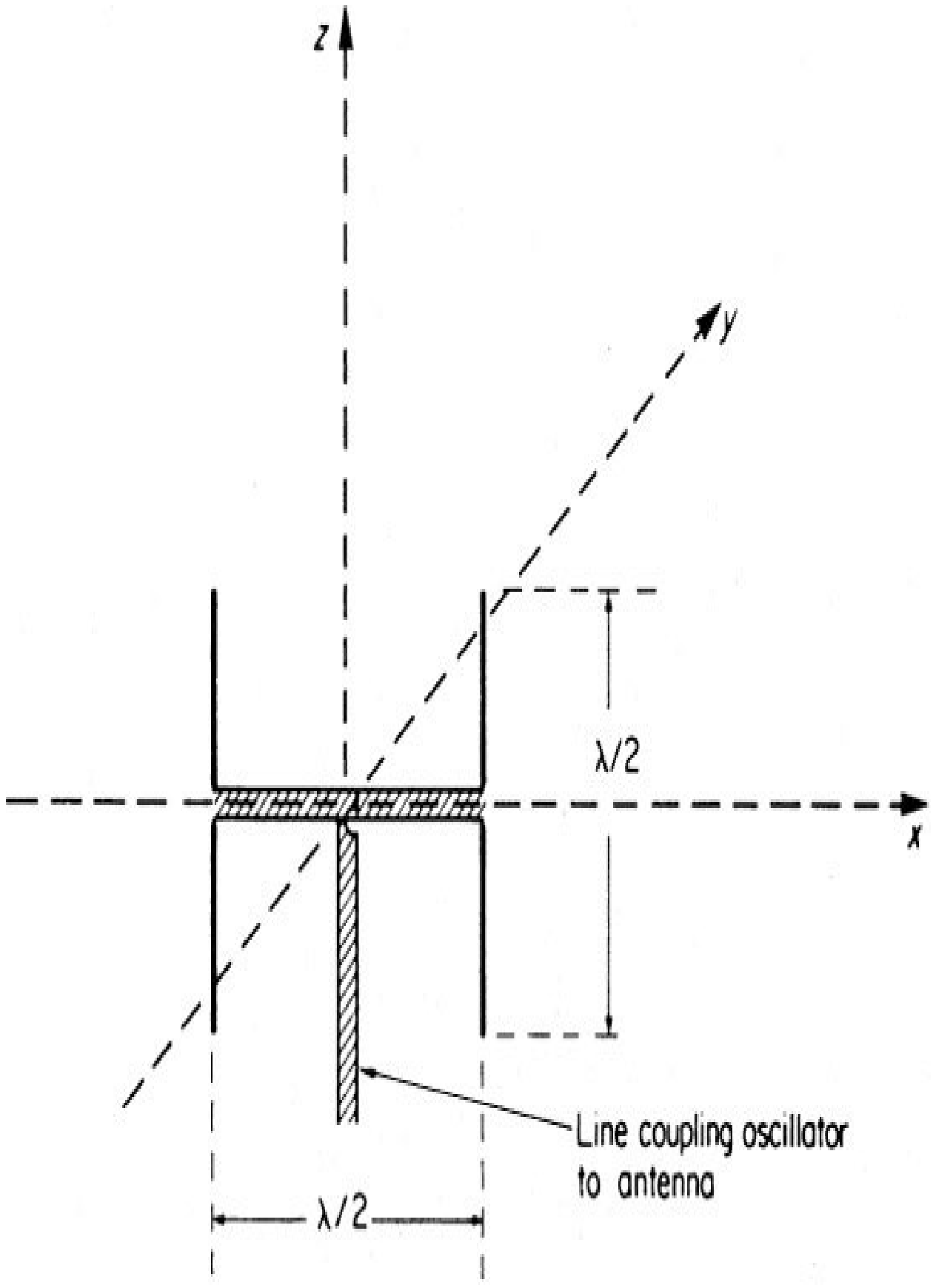
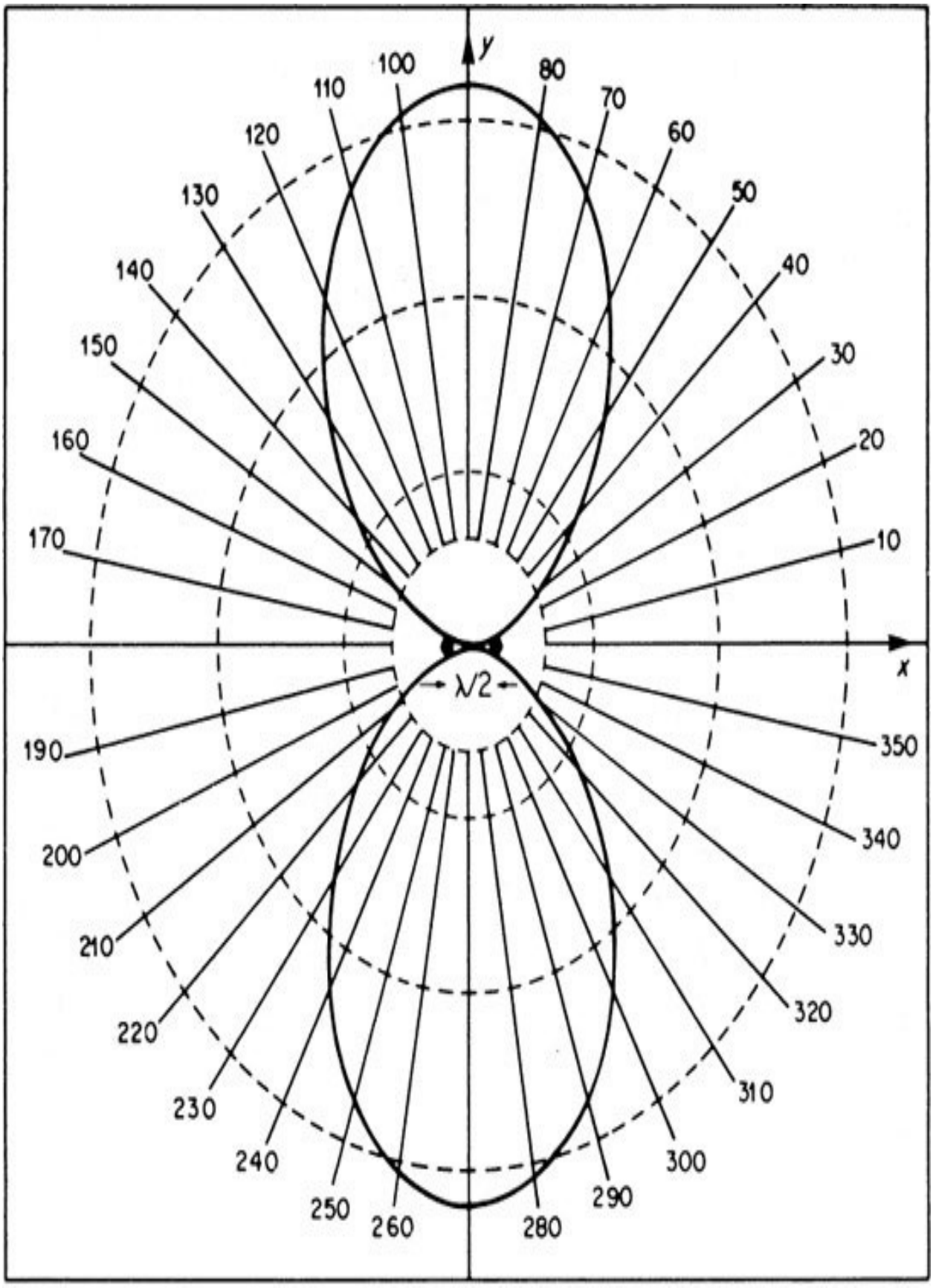

z
y
x
λ/2
λ/2
Line coupling oscillator
to antenna

Figure 13.9. The relative field strengths in the x–y plane far from the antenna shown in Figure 13.8. Note that this is not similar to the previous polar diagrams. This one gives the relative field strengths at points on a circle drawn in the x–y plane and centred on the antenna, and not at different points on the surface of a sphere. The fields now depend on azimuthal angle ϕ which is the angle shown on the scale, and a different diagram would have to be drawn to show the fields in another plane.


y
x
λ/2
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
190
200
210
220
230
240
250
260
280
290
300
310
320
330
340
350

PROBLEMS 13

13.1 A half-wave dipole transmitter in the centre of a city is required to give signals to cars travelling in the city within a radius of 10 km. If the minimum field amplitude required is 0.02 Vm^{-1} what must be the mean power of the transmitter? Why would the frequency chosen be likely to be around 100 MHz?

13.2 Give the impedance of a quarter wavelength length of coaxial line suitable for connecting a 50 Ω line to a 72 Ω aerial.

13.3 An antenna consists of four vertical half-wave dipoles separated from each other by one wavelength in the East–West direction. Calculate and sketch the horizontal polar diagram of the array when the dipoles are fed in phase with equal currents.

13.4 Why is an antenna much shorter than a wavelength inefficient as a radiator of radio waves? If a commercial radio station transmitting at a wavelength of 500 m did not wish to string a cable 250 m long well above ground to form a half-wave dipole antenna, what might it do to make an efficient aerial system?

13.5 Electric quadrupole radiation is produced when the currents in two very small dipoles, placed side by side very close to each other, oscillate 180° out of phase. Show that the electric radiation field produced is proportional to the cube of the angular frequency. Hence show that the probability per unit time for the decay of an excited microscopic energy level via quadrupole radiation is proportional to the fifth power of the photon energy.

13.6 Determine the ratio of the e.m.f.s set up in a short wire of length l and in a closed loop of wire of area S when they are used as receiving aerials. Consider the cases where both aerials are orientated so as to produce maximum e.m.f.s.

13.7 An oscillating current $I_0 \cos \omega t$ in a small loop of wire of area a^2 constitutes an oscillating magnetic dipole with maximum magnetic dipole moment $m_0 = I_0 a^2$. The mean power radiated by such a dipole is . This relation is obtained from Equation (13.23) by replacing the electric dipole moment d_0 by the magnetic dipole moment divided by the speed of light in vacuo m_0/c. Determine

the radiation resistance of the loop.

A small loop of area a^2 is connected to a matched load. Radiation of angular frequency ω is incident from a direction in the plane of the loop and is plane polarized with its magnetic field vector perpendicular to the plane of the loop. Find the ratio of the mean power scattered and absorbed by the loop to the mean value of the Poynting vector in the incident wave.

The ratio calculated above has the dimensions of a length squared and is often referred to as the cross-section for combined scattering and absorption.

* See section 14.8 for a rigorous discussion.

* See for example section 11.7 in Optics, Second edition, by F. G. Smith and J. H. Thompson, J. Wiley, London, 1988.

* Ibid., section 9.4.

CHAPTER 14

Electromagnetism and special relativity

14.1 INTRODUCTORY REMARKS

Maxwell's equations were used in section 11.1 to predict that light (or other electromagnetic waves outside the visible spectrum) travels in vacuo at the constant speed . The derivation of this result made no reference to the velocity of the source emitting the light. Experiment confirms that an observer always finds the value c for the speed of light, whatever may be the velocity of the light source relative to a frame of reference fixed in his laboratory (i.e. the velocity measured in a set of coordinates fixed in his laboratory). For example, the speed of the radiation emitted in the decay of π-mesons is measured to be c, even if the mesons themselves have been moving through the laboratory at speeds close to c*. Furthermore, there is no reason to suppose that the constancy of the speed of light is a peculiarity of our laboratory frame of reference—one would expect to find the same result in a space station or in a laboratory on another planet.

Questions about observations in different frames of reference are the concern of the theory of relativity. The special theory of relativity is restricted to statements about inertial frames of reference. Inertial frames are those in which the law of inertia holds, i.e. those in which all bodies move with a constant velocity unless they are subjected to external forces. All inertial frames move with a constant velocity relative to one another. The special theory accepts that there is no way to pick out one of these frames as having a more fundamental importance than any of the others, and treats all inertial frames on an equal footing. If inertial frames are to be treated alike, then it must be possible to find a way of expressing physical laws which is common to all of them.

The equivalence of inertial frames and the constancy of the speed of light are the two ideas forming the basis of the special theory of relativity: they are the axioms of the theory. Written out in full, the axioms are

Axiom No. 1 Physical laws must have the same form in all inertial frames of reference.

Axiom No. 2 All observers find the same value c for the speed of light.

The results of the special theory of relativity follow from the requirement that the laws of physics should be in accordance with these two axioms. In mechanics, for example, the equivalence of mass and energy, and the increased mass of bodies moving at speeds close to c, follow from the relativistic laws proposed by Einstein. These effects are both well established experimentally, and constitute a powerful corroboration of the theory.

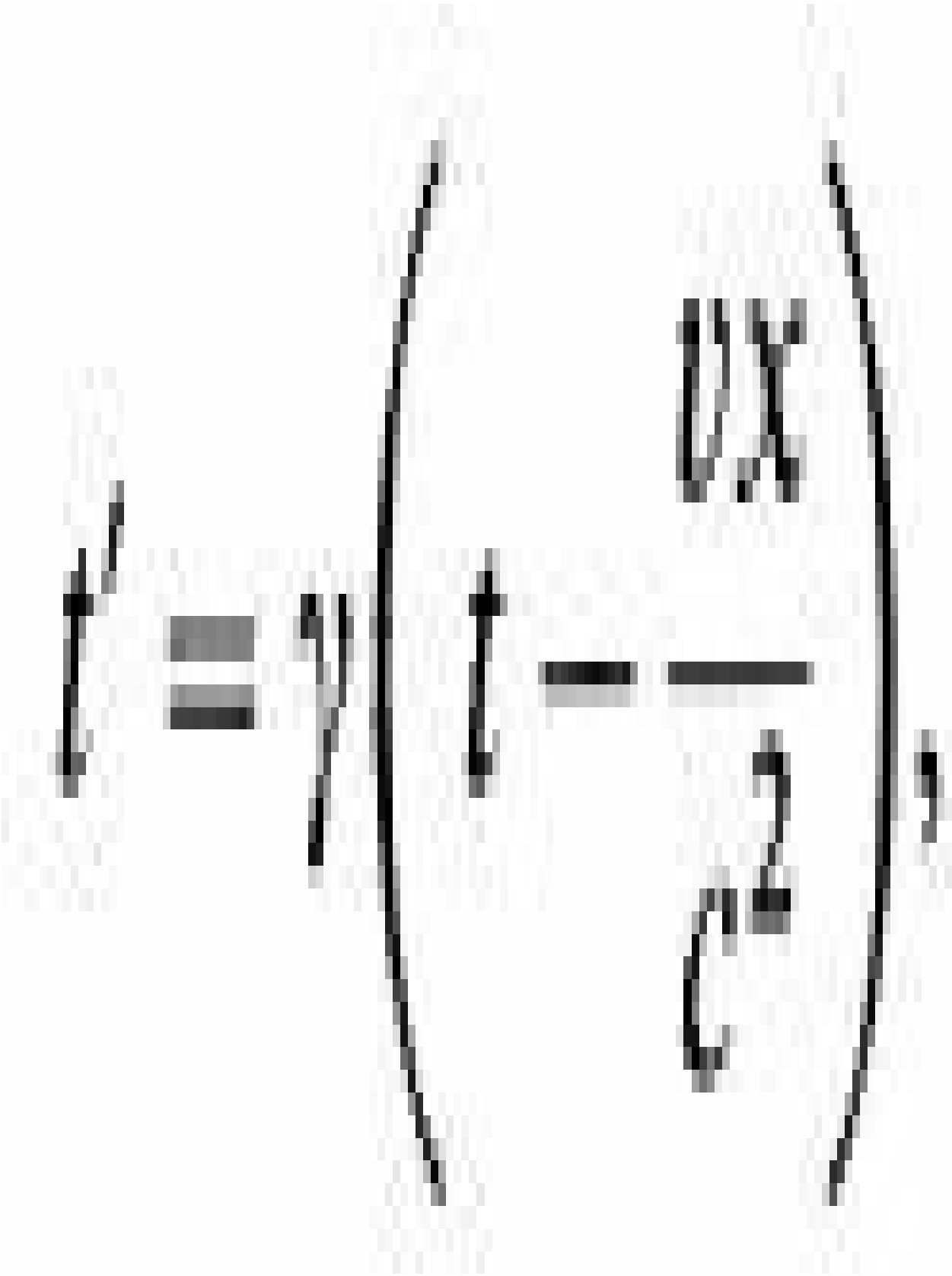
The purpose of this chapter is to test the laws of electromagnetism against the axioms of special relativity. We shall show that although observers in different frames of reference disagree about the magnitude of electric and magnetic fields, they all agree that the fields obey Maxwell's equations. In other words the laws of electromagnetism, as embodied in Maxwell's equations, are in accordance with the first axiom. Now from Maxwell's equations the speed of light is deduced to be c, and the second requirement of the theory is automatically fulfilled. The laws of electromagnetism therefore need no modification in moving frames of reference: they are already fully consistent with the special theory of relativity.

There is another way of looking at the matching of electromagnetism and special relativity. Not only is relativity a well-established theory, but its axioms are of such a general nature (axiom No. 1 seems almost intuitive) that we should like to place the theory very high in the hierarchy of physics. Other laws must be made to fit in with the special theory of relativity. In this spirit, we start in section 14.7 with the Coulomb field of a stationary point charge, and use relativistic arguments to acquire information about the fields generated by moving charges. In effect Maxwell's equations are derived from Coulomb's law, so that the whole of electromagnetism is seen to follow naturally from electrostatics.

14.2 THE LORENTZ TRANSFORMATION

Before discussing electromagnetic fields, we shall remind the reader of the Lorentz transformation relating space and time coordinates in different frames of reference. Consider two inertial frames S and S′ having a common origin at t = 0, but with the origin of S′ moving along the x-axis of S at a speed v (see Figure 14.1). An event occurring at time t in the frame S at the point with coordinates (x, y, z) is seen in S′ as occurring at time t′ at the point (x′, y′, z′). The relation between the two sets of coordinates is given by the Lorentz transformation, which follows from the postulates of special relativity. The relation is

(14.1)



where

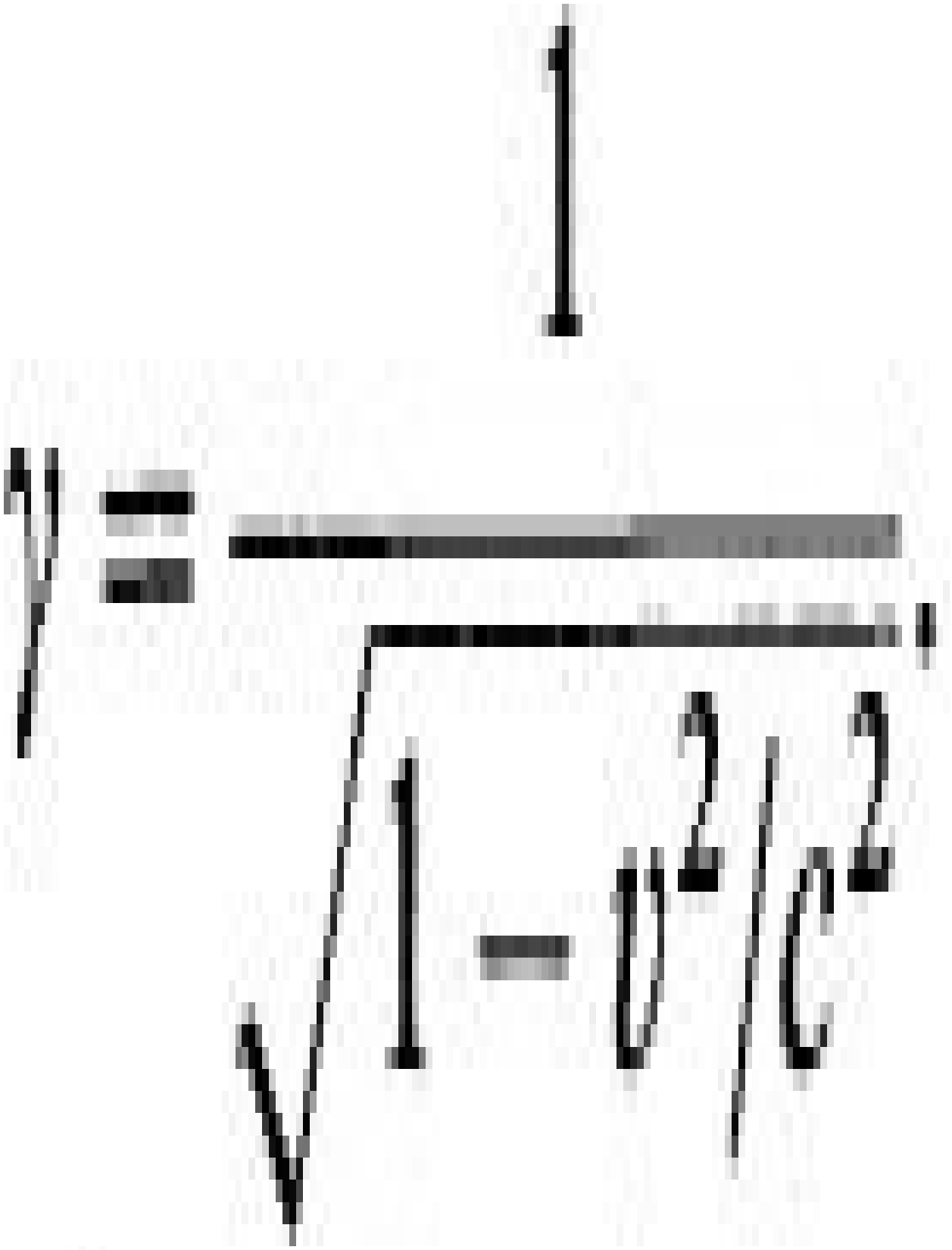
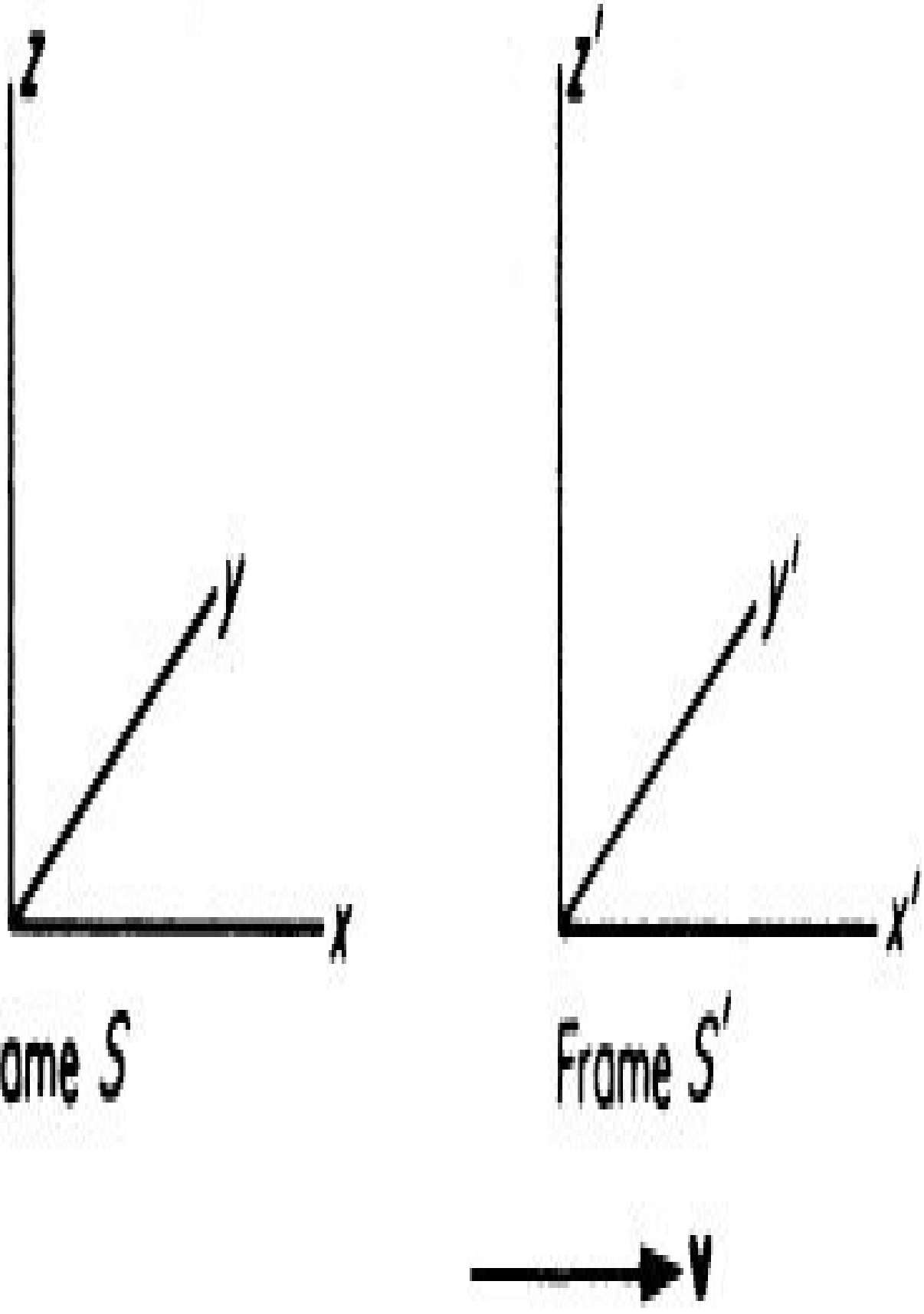
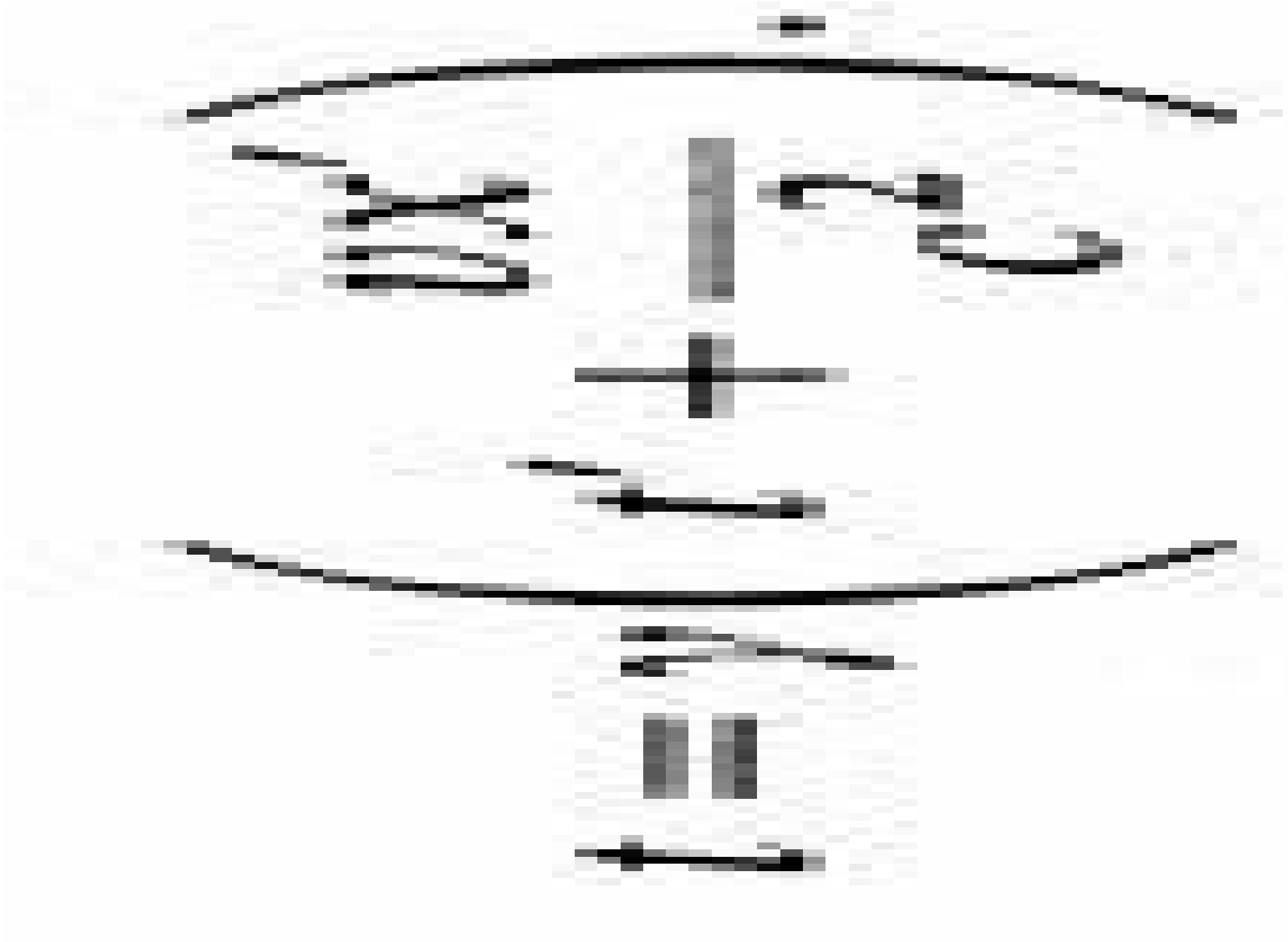


Figure 14.1. Two inertiaf frames of reference in relative motion.


z
y
x
z'
y'
x'
Frame S
Frame S'
v

For an observer in S′, the origin of S appears to be moving along the x′-axis at a speed (− v). The two observers are on an equal footing, and the inverse Lorentz transformation, giving coordinates in S in terms of their values in S′, is obtained simply by replacing v by (− v) in Equation (14.1):

(14.2)



(The consistency of the two sets of equations can be checked by substituting 14.2 into (14.1), leading to $x' = x'$ etc.)

Two features of the Lorentz transformation will be particularly prominent when we go on to electric and magnetic fields. These are the FitzGerald contraction, and time dilatation, which refer to changes in the scales of length and time respectively, for observers in the frame S and S′. First, time dilatation. Suppose that a clock at a fixed point x' in the frame S′ measures a time interval $(t'2 - t'1)$ between events occurring at x' at the times $t'1$ and $t'2$. For an observer in S, these events appear to occur at times given by Equation (14.2) as

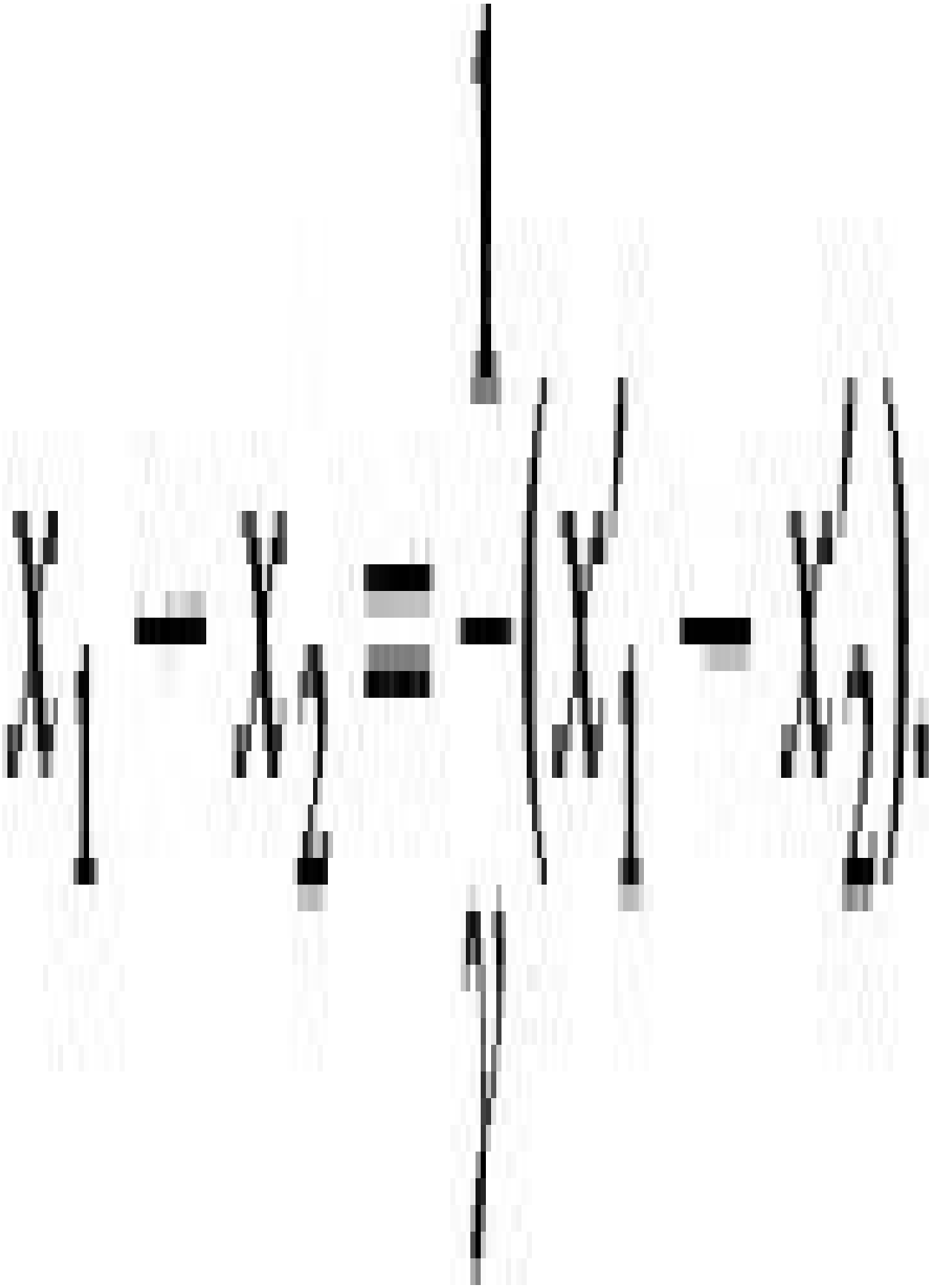
$$t_1 = \gamma\left(t'_1 + \frac{vx'}{c^2}\right)$$

$$t_2 = \gamma\left(t'_2 + \frac{vx'}{c^2}\right).$$

He thus considers the time interval to be $(t2 - t1) = \gamma(t'2 - t'1)$. This is larger than the time $(t'2 - t'1)$ measured by the clock in S′ since is greater than unity. In other words, the moving clock appears to run slow.

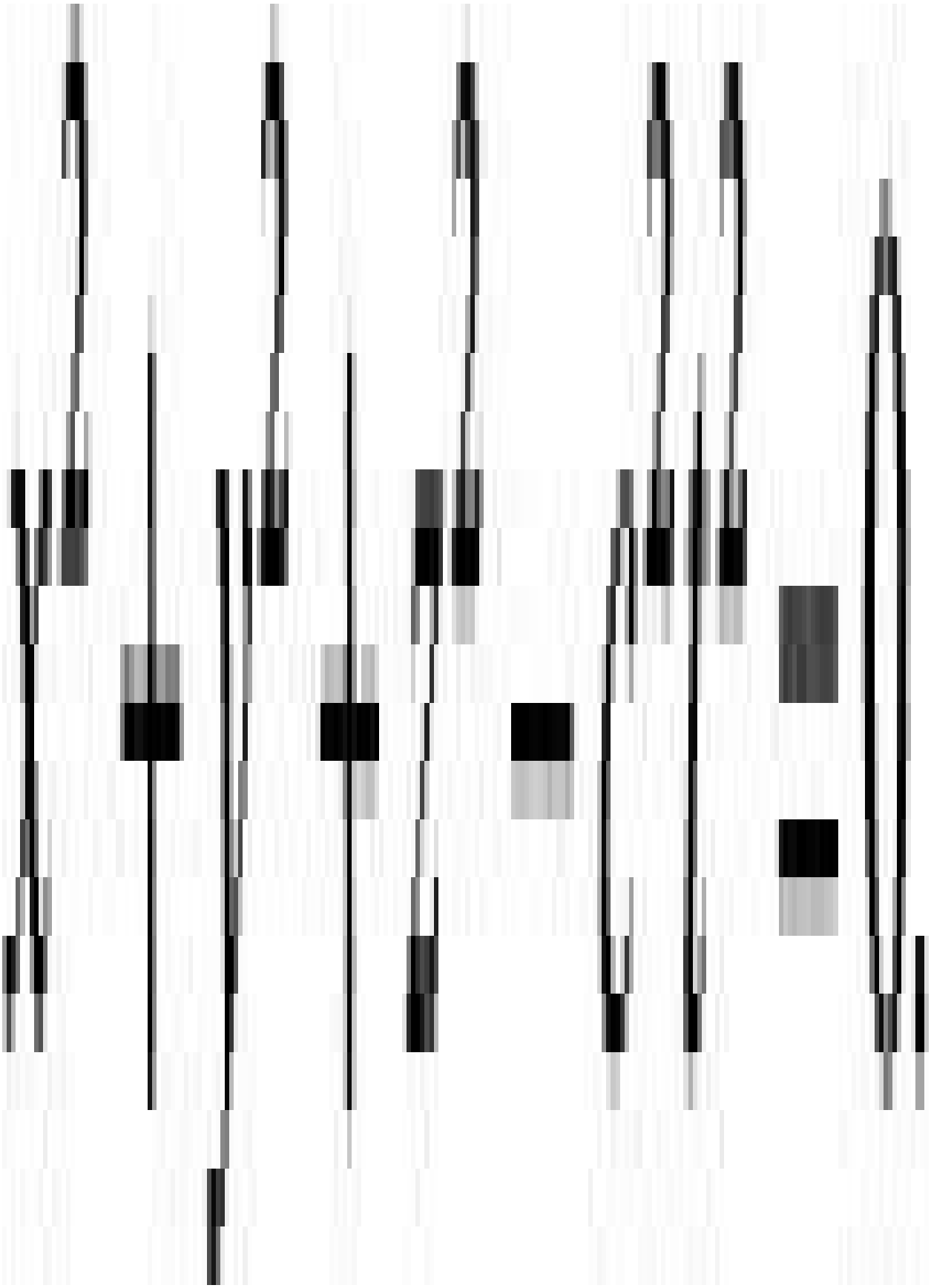
We will now examine the analogous problem of the apparent length of a moving rigid rod. Imagine a rigid body, with ends which are fixed in S′ at the points $x'1$ and $x'2$. At a definite time t in the frame S, the ends are seen to be at x1 and x2, where from (14.1)

Now the length of the moving body as measured in S is the distance between its ends at the same time t, i.e.

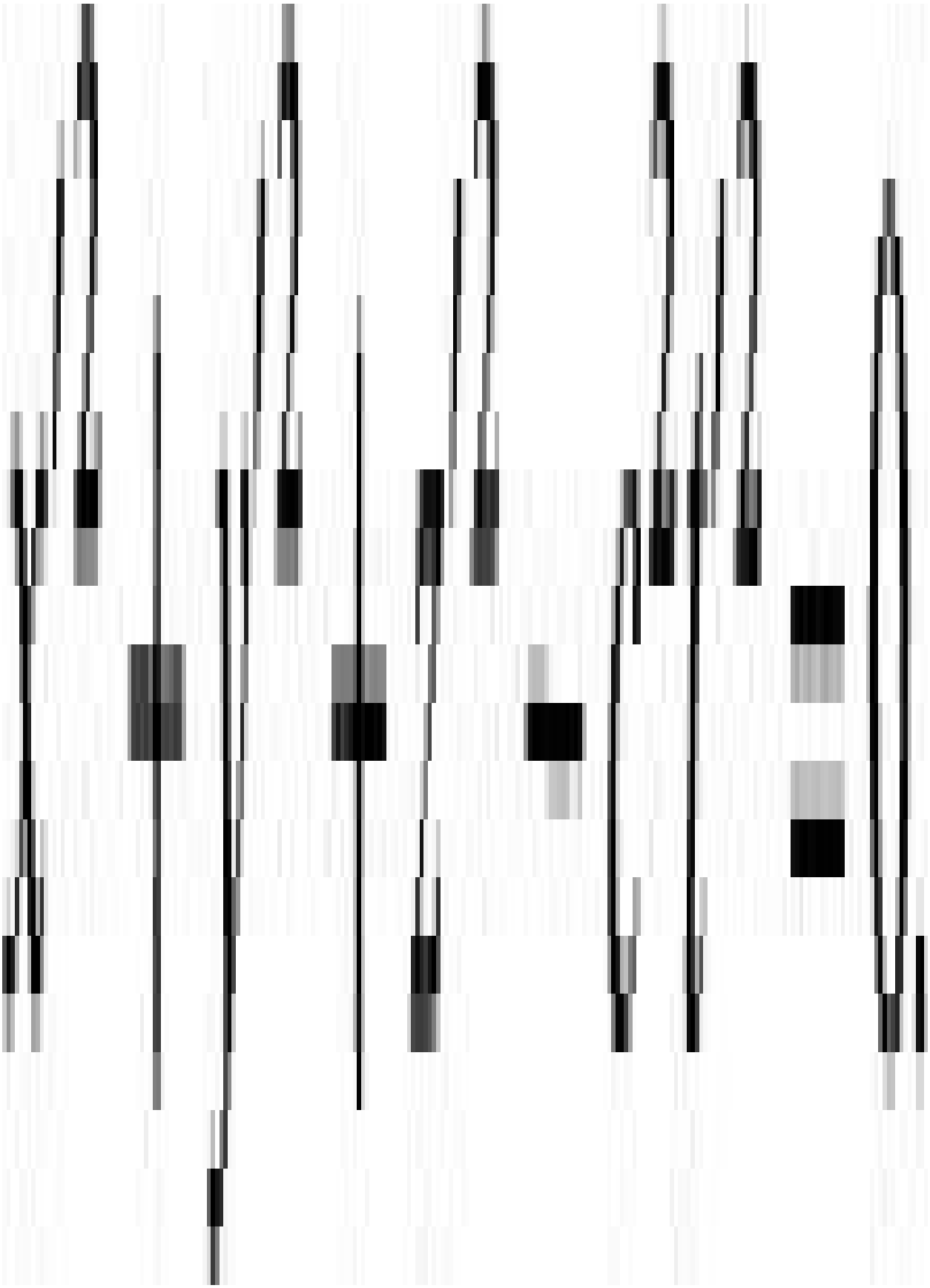


The moving body appears to be shorter, its length having contracted by the factor 1/γ. This effect is called the FitzGerald contraction. Notice that the FitzGerald contraction only operates along the direction of motion: since y′ = y and z′ = z, the apparent dimensions of the body are unaltered in directions perpendicular to its motion.

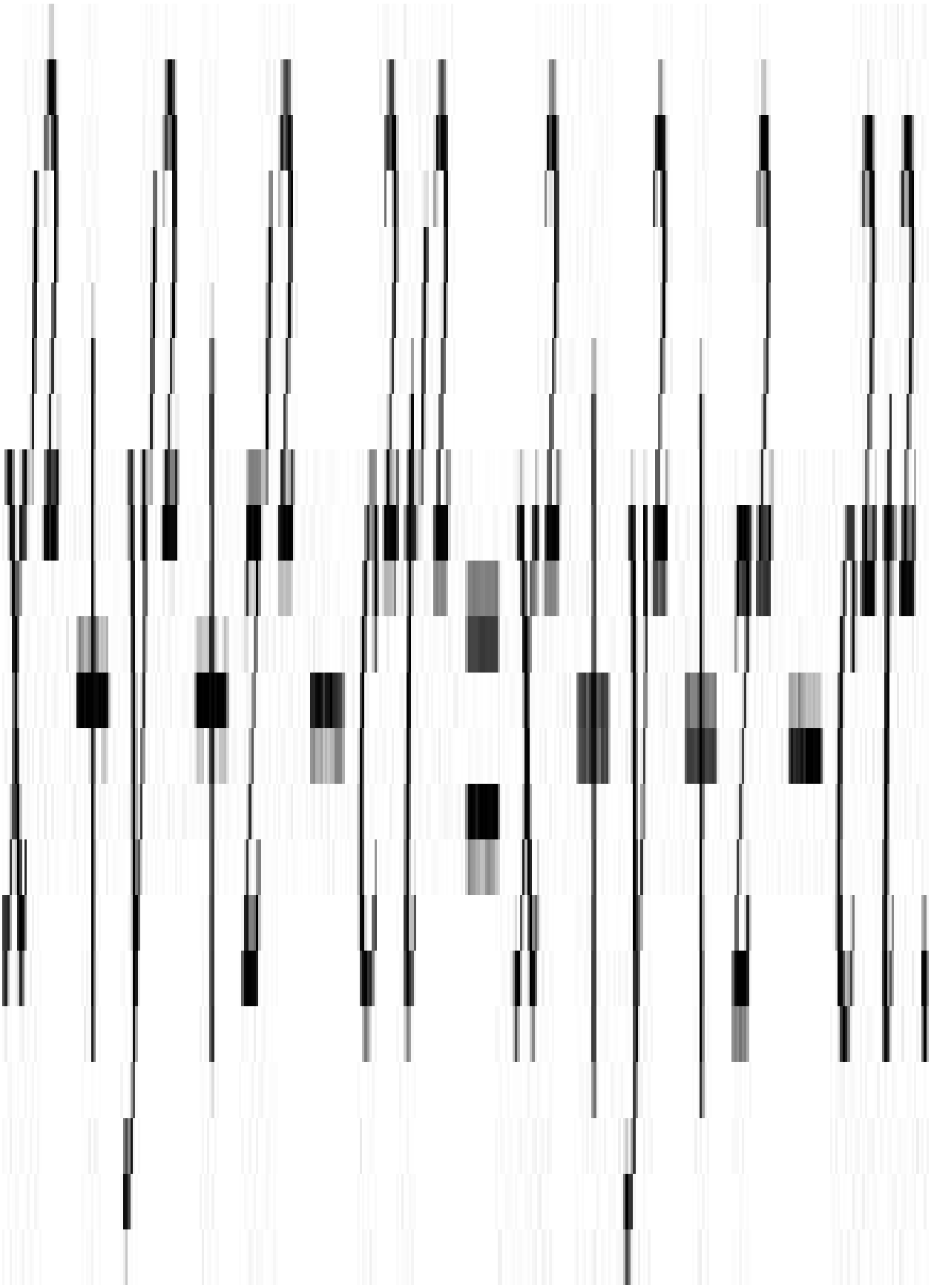
Although distances and time intervals are changed by the Lorentz transformation, observers in S and S′ must still agree that light waves travel at a speed c. If a light signal leaves the origin at time zero and propagates as a spherical wave, an observer in S describes the locus of the spherical wave-front at time t by the equation



Similarly, as seen by an observer in S′ the locus at time t′ is given by



These two equations are consistent with the Lorentz transformation, since we find by substituting from Equations (14.1) that



A quantity like $(x^2 + y^2 + z^2 - c^2t^2)$ which always has the same value for observers in S and S′ is said to be invariant under the Lorentz transformation, or to be Lorentz invariant.

14.3 CHARGES AND FIELDS AS SEEN BY DIFFERENT OBSERVERS

Charge

An atom carries negative charges in its electron cloud, and positive charges on the protons in its nucleus. Both protons and electrons are moving about within the atom, yet their positive and negative charges exactly cancel so that the atom as a whole preserves strict electrical neutrality. This implies that the charge on each elementary particle appears to be ± e whether or not the particle is moving relative to the observer. In other words, the electronic charge is invariant under Lorentz transformations. On a charged body containing many atoms, different observers could in principle count up the number of neutral atoms and the number of spare elementary charges left over. Since the observers agree that each spare elementary charge has a magnitude + e or − e, they would also agree about the net charge on the body. It follows that the net charge carried by any macroscopic body is Lorentz invariant.

Charge density of a line of equally spaced charges

Charge density, unlike total charge, is not Lorentz invariant. Imagine a series of positive charges each of magnitude + e, strung out along a line with equal spacings as shown in Figure 14.2. Suppose that an observer in a frame of reference S is at rest relative to the charges, and that in S the N charges labelled 1 to N span a distance of unit length. Then this observer sees a linear charge density Ne coulombs/m. The observer in S′ moving with a speed v parallel to the line of charges agrees that each of the N charges carries a charge + e, but he does not agree that they cover a unit of length. Because of the Lorentz contraction, in his view the N charges have been squashed up into a distance $1/\gamma$, and the linear

charge density is

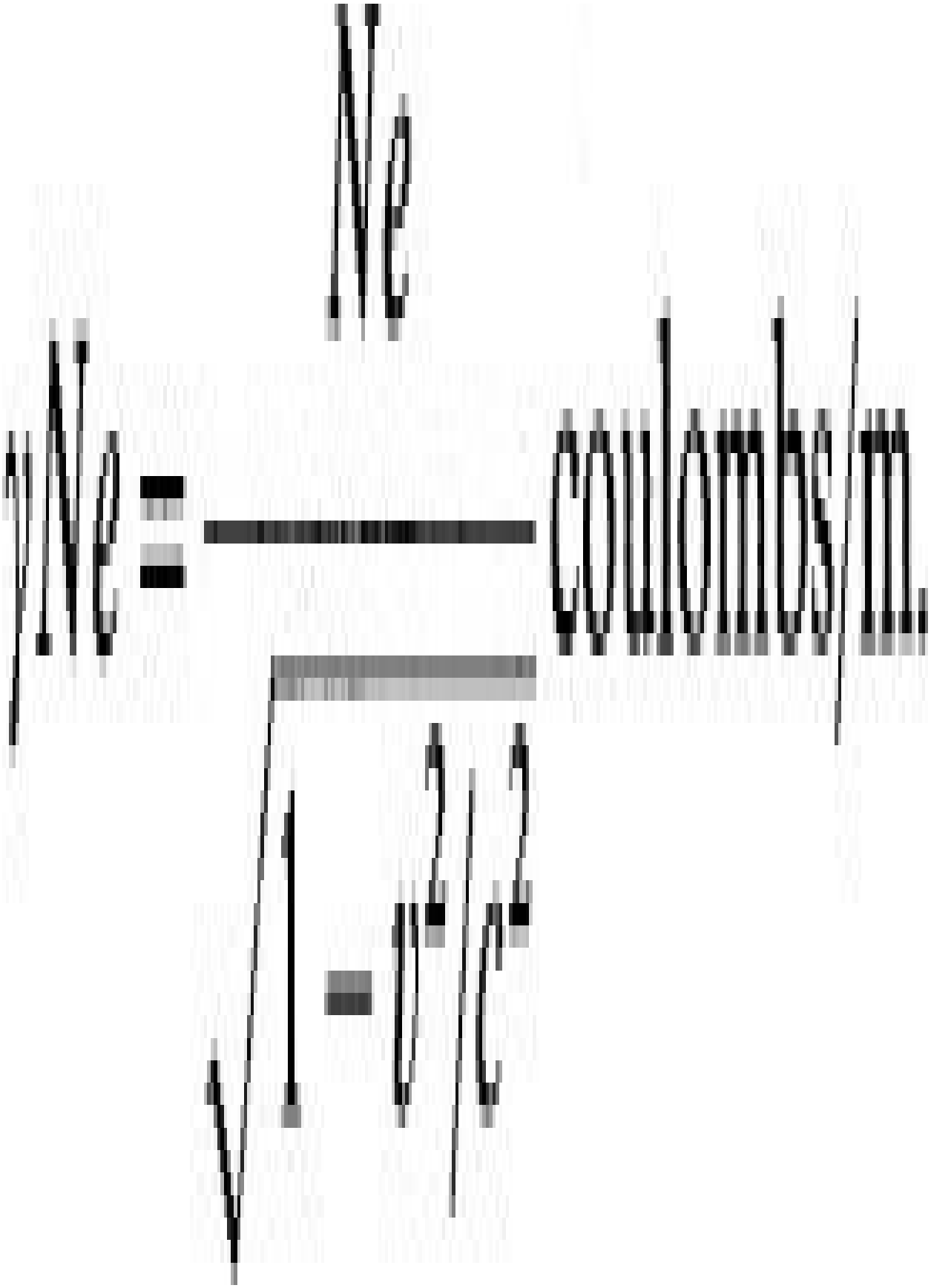
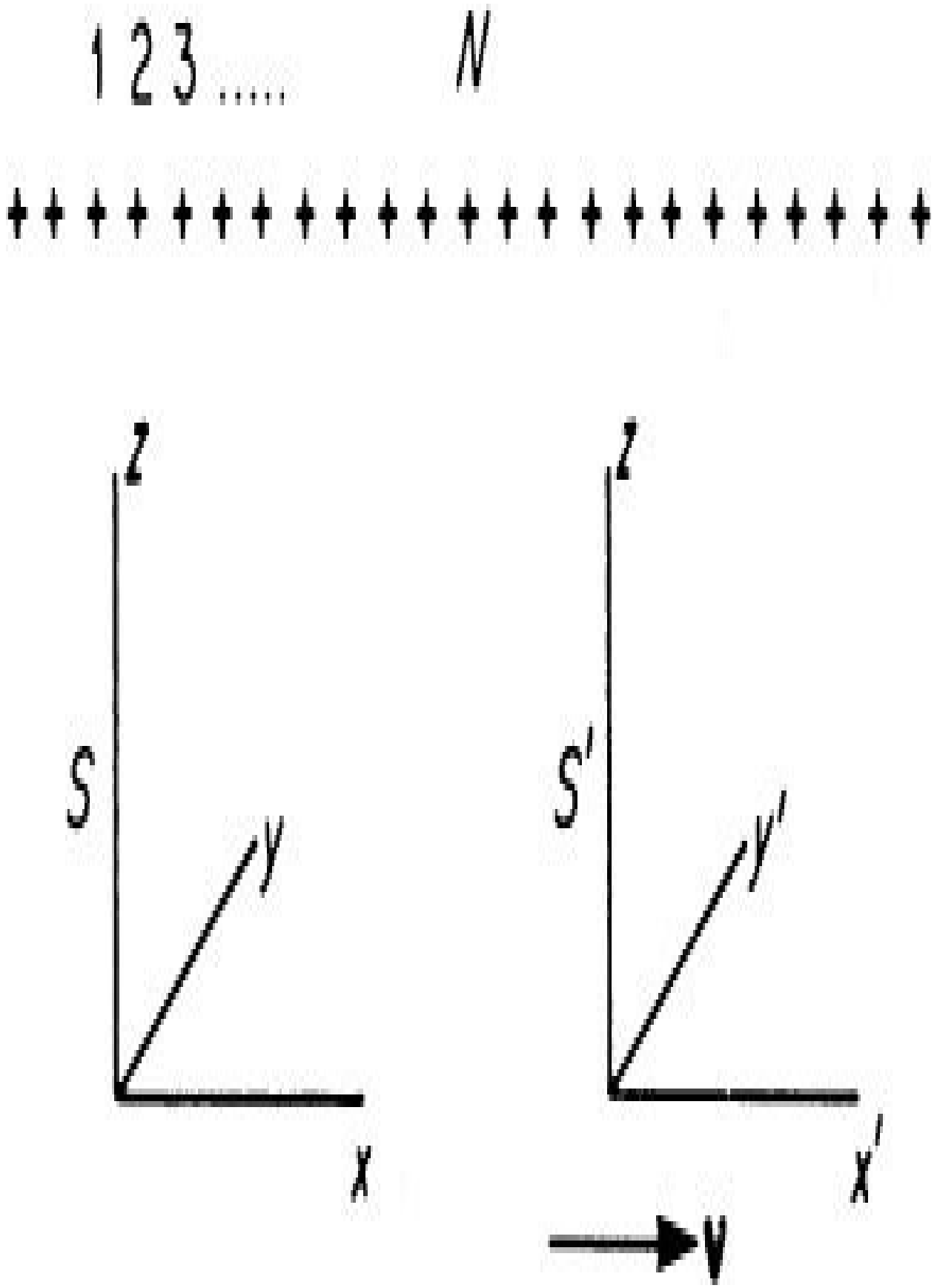


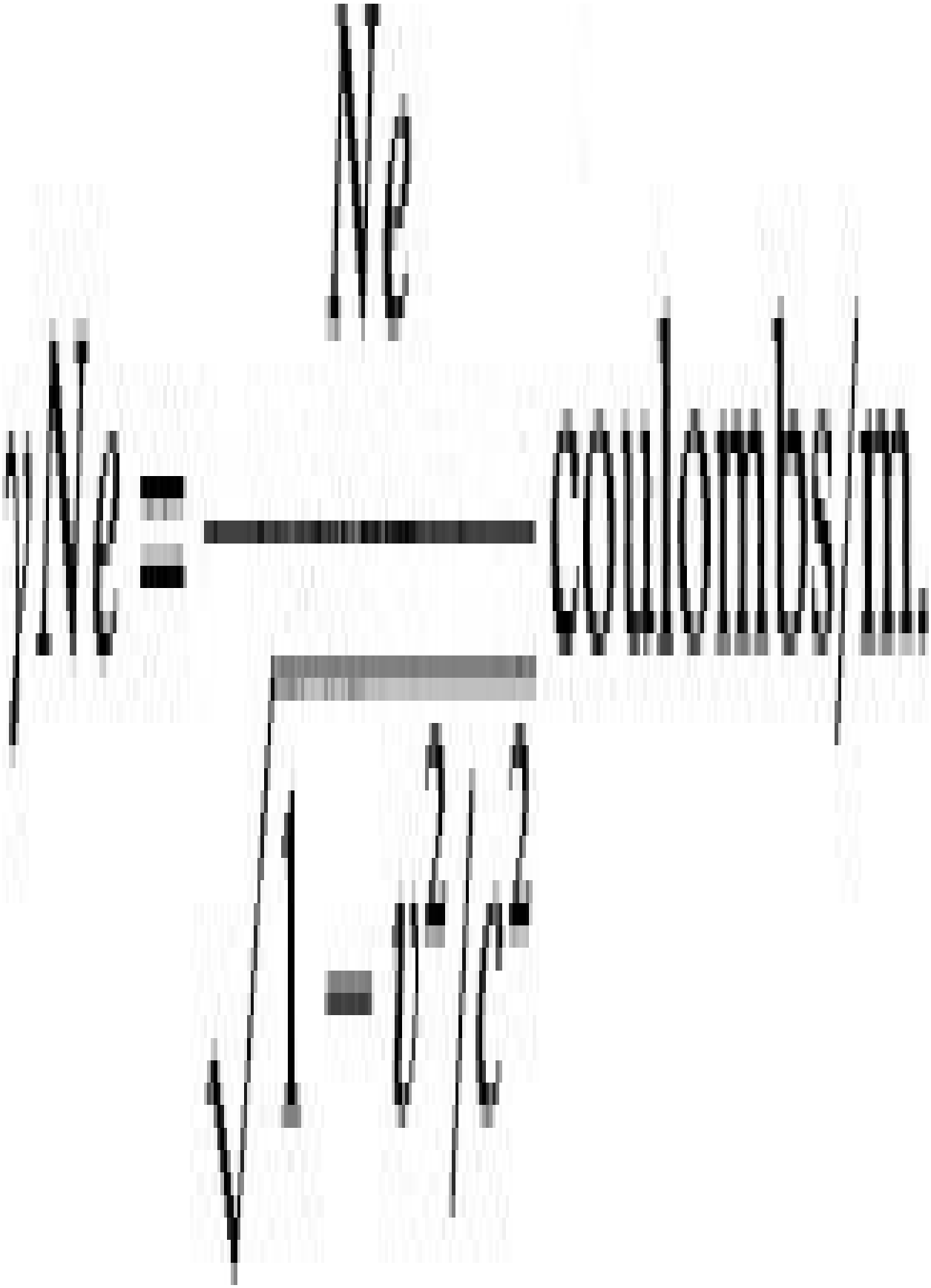
Figure 14.2. A line of charges which is stationary in S appears to be moving from right to left in S′.


1 2 3 .....
N
z
S
y
x
z
S'
y'
x'
V

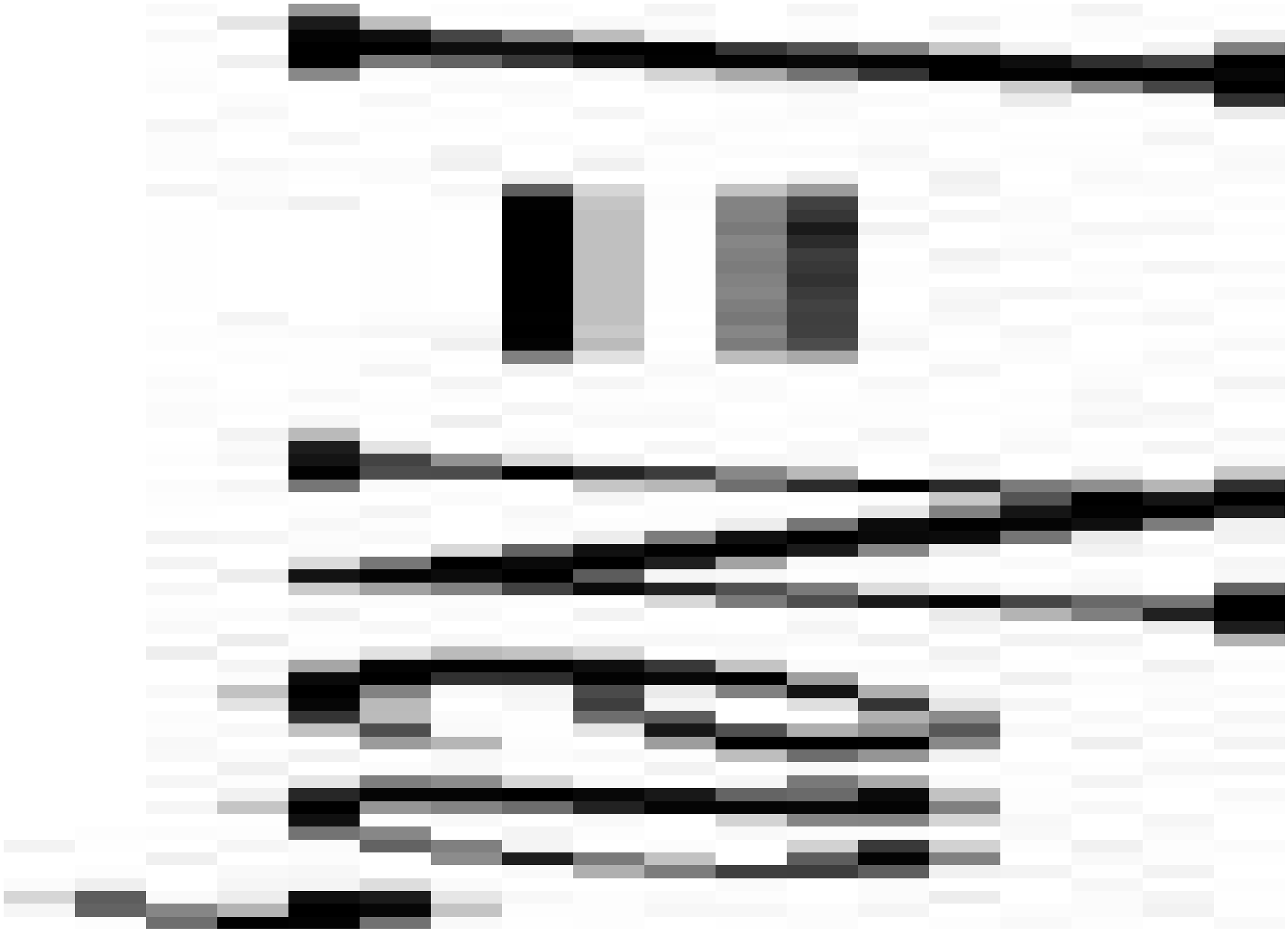
Electric and magnetic forces

Let us now investigate the forces exerted by electric and magnetic fields, as viewed by observers in S and S′. Suppose that in the frame S, a wire which has no net charge carries a current in the x-direction. A charge + q moving parallel to the wire at a speed v experiences a magnetic force pushing it towards the wire. But from the point of view of the observer in a frame S′, in which this charge is at rest, it cannot be acted upon by any magnetic forces! Yet he must agree with the observer in S that there is a force on q, pointing towards the current-carrying wire. What does he consider to be the origin of this force? The answer is that the observer in S′ does not think that the wire is electrically neutral; he says that there is an electrostatic field around the wire. As we shall see, the difference arises because of the FitzGerald contraction.

For an observer in S, the neutral wire contains equal numbers of positive and negative charges ± e, each with a density N charges per unit length, say. Assume that all the negative charges are at rest, and that all the positive charges are moving at the same speed v (the calculations are made easier by this assumption, although the final result does not depend on it). The current flowing along the wire is



and at a distance r from the wire the observer in S sees a magnetic field of magnitude (see Equation (4.37))

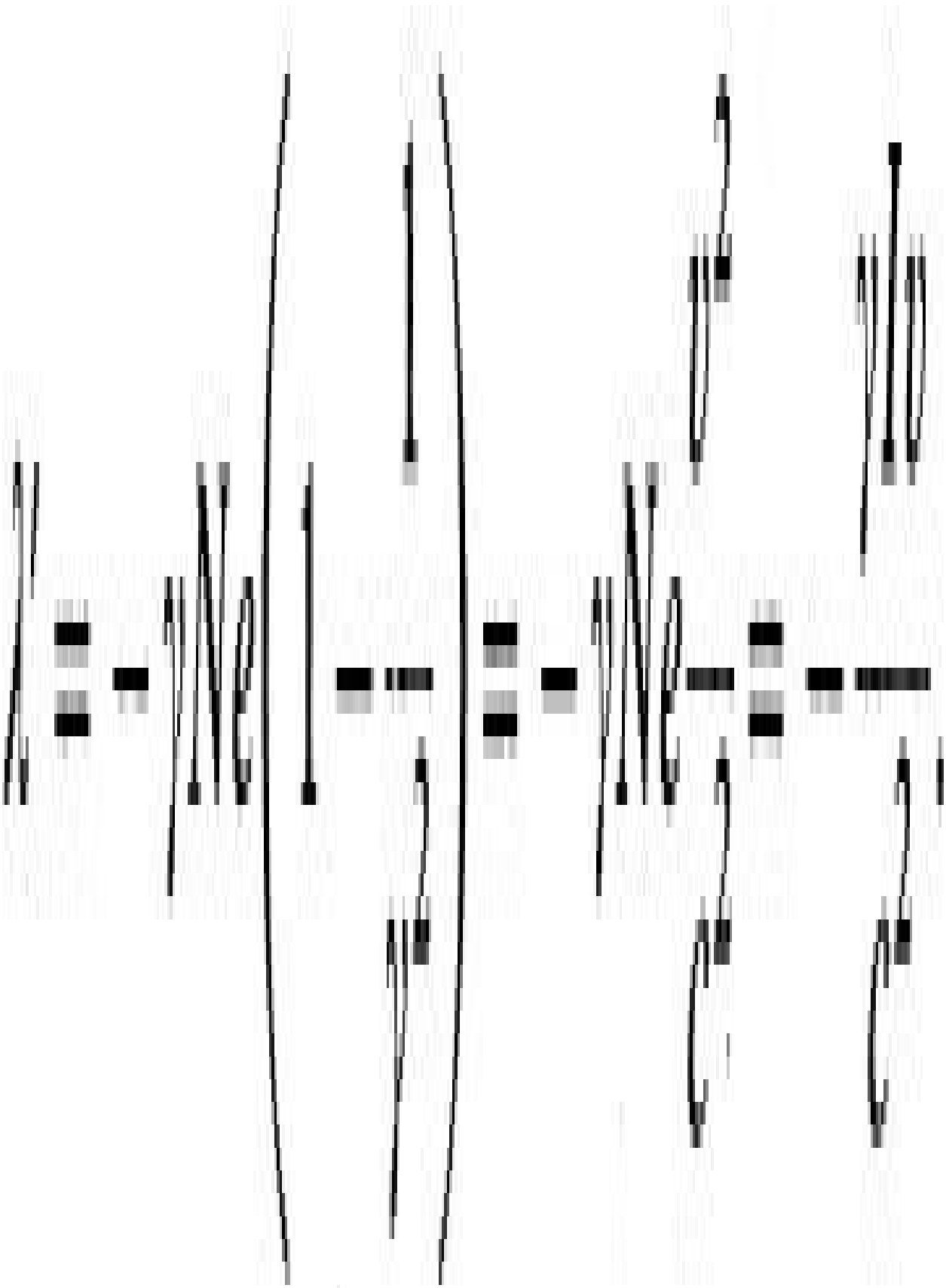


The charge + q moves at a speed v in the x-direction in S, perpendicularly to the magnetic field, and if its distance from the wire is r it experiences a force

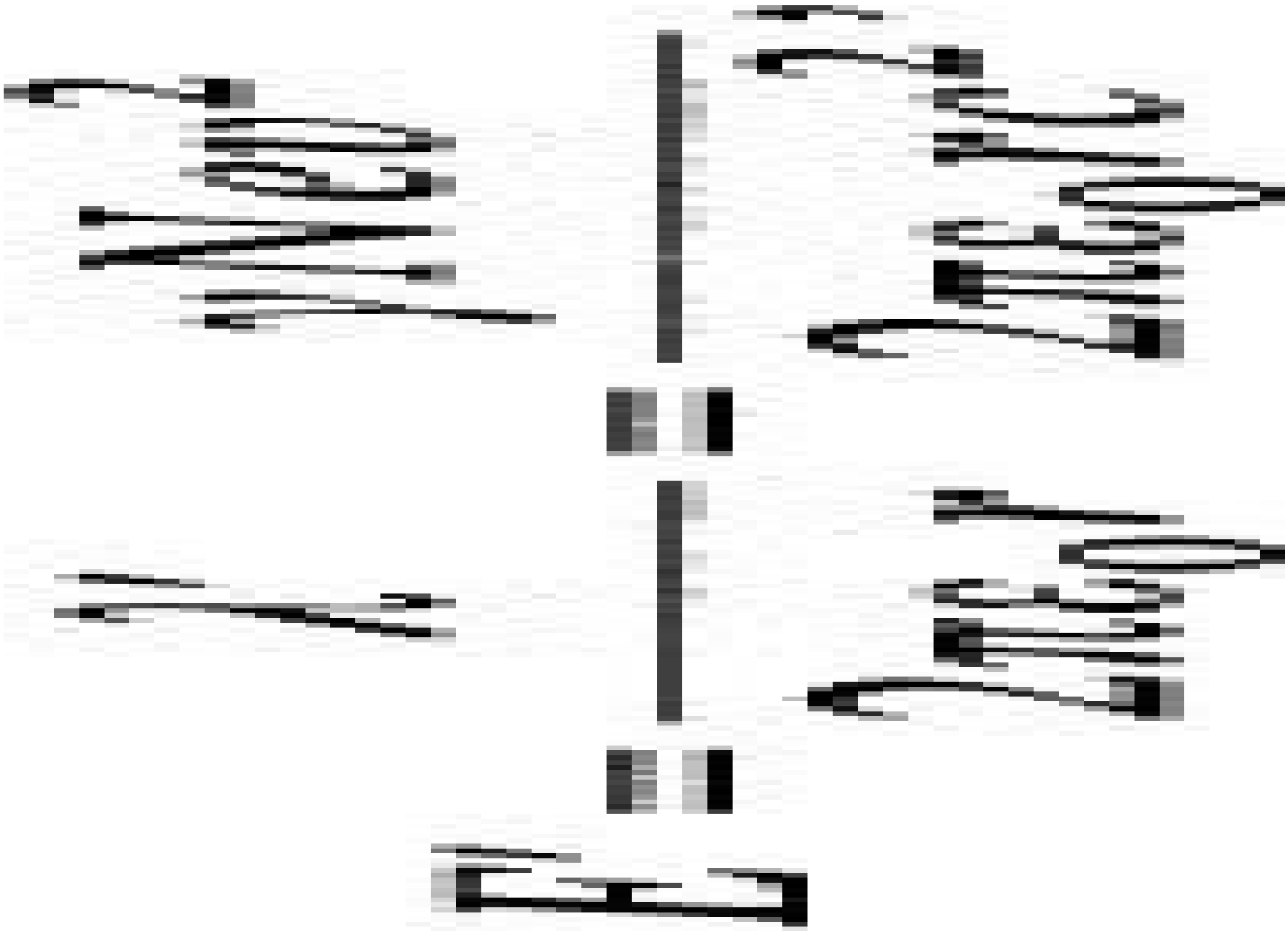
(14.3)

Now let us evaluate the force from the point of view of the observer in S′. The negative charges which were at rest in S are moving at a speed v relative to him. They are squashed up by the FitzGerald contraction, and he measures their linear charge density to be (− γ Ne). On the other hand the positive charges are at rest in his frame; the contraction works the other way round, and he measures a diminished density Ne/γ. From his point of view the wire has a net linear charge density*

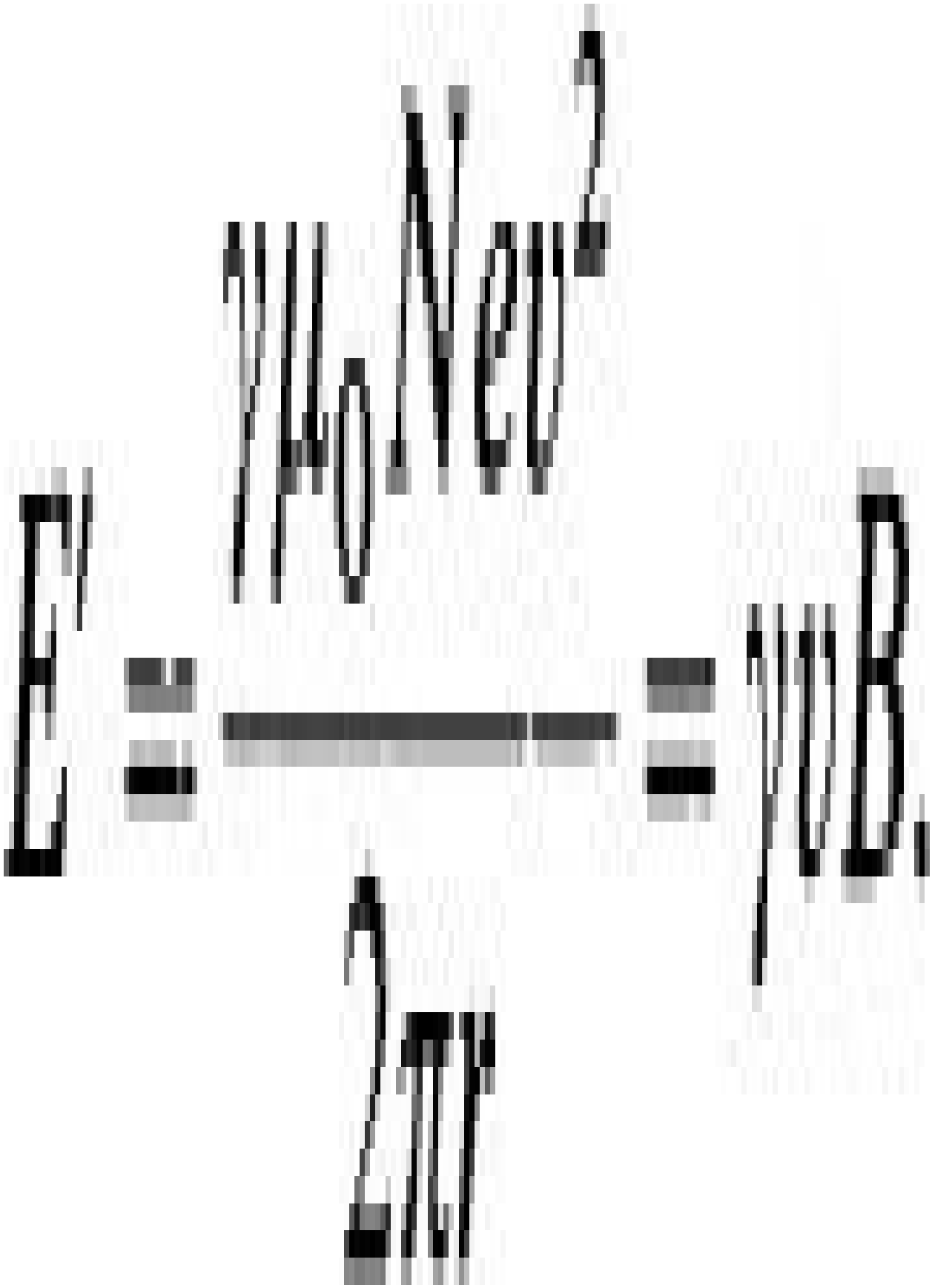
(14.4)



Distances perpendicular to the motion are not affected by the Lorentz transformation, and in S′ the stationary charge + q is still at a distance r from the wire. It experiences an electric field pointing towards the wire, of magnitude



or, replacing $\varepsilon_0 c^2$ by $1/\mu_0$,

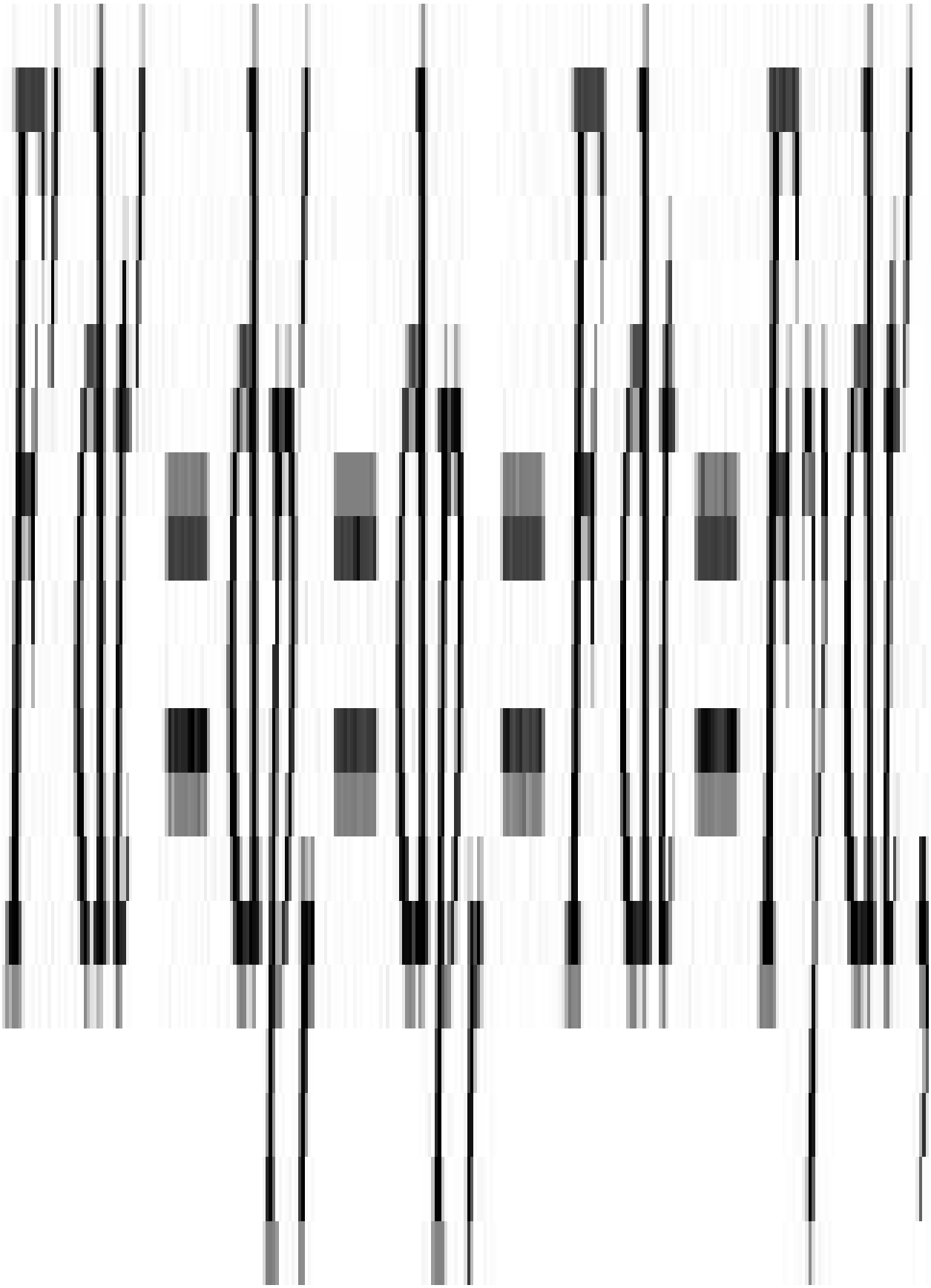


The electrostatic force on the charge is

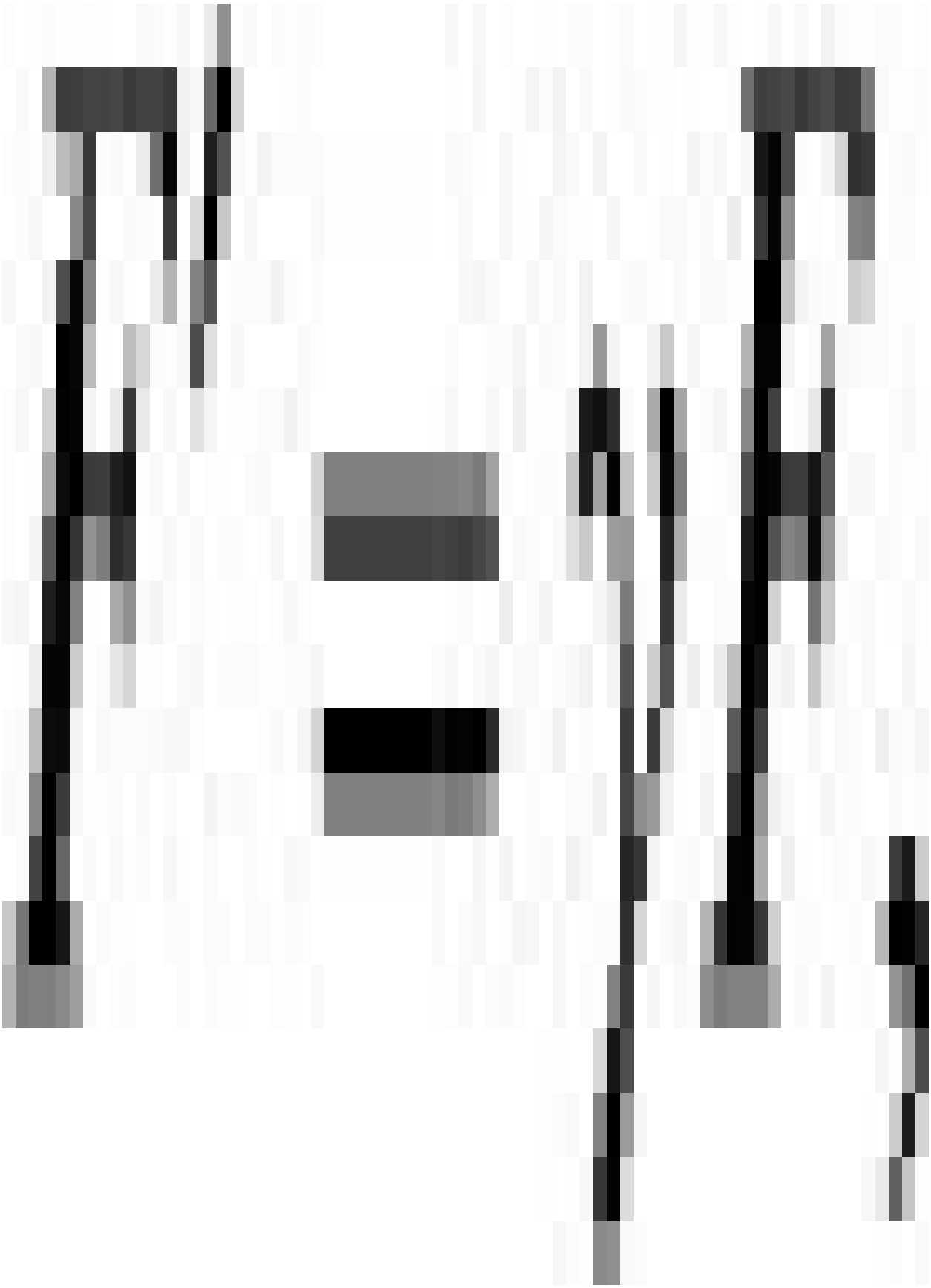
(14.5)

by comparison with Equation (14.3).

The force which was purely magnetic in the frame S appears to be purely electrostatic in S′. Observers in the two frames disagree not only about the origin of the force but also about its magnitude, the electrostatic force in S′ being larger by the factor γ. The factor γ arises because of time dilation, as we shall see by considering momentum changes in the frames S and S′. In S′ the force F′ gives q a component of momentum p′r towards the wire such that F′ = dp′r/dt′. In a time dt′ the charge thus acquires a momentum dp′r = F′ dt′. Now transverse momentum is not altered by a Lorentz transformation; during the time measured in S′ as dt′, an observer in S agrees that q acquires a transverse momentum dpr = dp′r. But the observer in S thinks that a clock which is at rest relative to q goes slow, and he reckons that the time taken to acquire the momentum dpr is dt = γ dt′. Hence



and



in agreement with Equation (14.5)

For simplicity we have illustrated the interchangeability of electric and magnetic fields by imagining a current composed of a set of charges all travelling at the same speed as the moving charge q. But the argument applies more or less unaltered however the current is constituted. An observer moving at a speed v parallel to the direction of a current I in a wire always considers that the wire has a line charge density λ' given by Equation (14.4): $\lambda' = -\gamma Iv/c^2$. In ordinary conductors the mean drift velocity of the moving charge carriers is very small—about 10^{-4} m/sec for a current of one amp flowing in a wire of area 1 mm². If the line density of moving charge carriers in the wire is κ, then and $\lambda' =$. The charge density of the wire as seen by the moving observer is thus extremely small compared to the density of moving charge carriers.

14.4 FOUR-VECTORS

According to the special theory of relativity there is no absolute frame of reference; all coordinate systems in uniform relative motion are equally acceptable, and physical laws must take the same form in all of them. How can one test whether or not a particular law is in accordance with this requirement? Let us illustrate the kind of argument used to make such a test by considering first of all two fixed coordinate systems which have the same origin but whose axes point in different directions. For example, if the frame of reference S in Figure 14.3 is rotated through 90° about the z-axis, it coincides with the frame S′ The coordinates of the point P are not the same in S and S′. Rotation of the axes mixes up the x- and y-coordinates so that if P is at (x, y, z) in S, its coordinates (x ′, y′, z′) in S′ are given by x′ = y, y′ = − x, z′ = z. Any rotation of the axes leads to a linear transformation of this kind, in which the new coordinates are linear combinations of the old ones.

Because the orientation of the axes is of no consequence, it is often a good idea to use the vector notation. In specifying the point P by a position vector r, no reference at all is made to the orientation of the axes; r represents the same point in space whatever set of axes is chosen. There is a further advantage in using the vector notation. Other quantities besides position are vectors, and when the axes are rotated, the components of these other vectors transform in exactly the same way as the components of a position vector. For example, looking again at the frames S and S′ in Figure 14.3, if a momentum vector p has components (px, py, pz) in S, then its components in S′ are p′x = py, p′y = − px, p′z = pz. Because transformations of this kind are always linear, the sum of any two vectors will also transform in the same way when the axes are rotated. Now look at a vector equation such as the equation defining the force F acting on a charge q in fields E and B (Equation (4.16)):

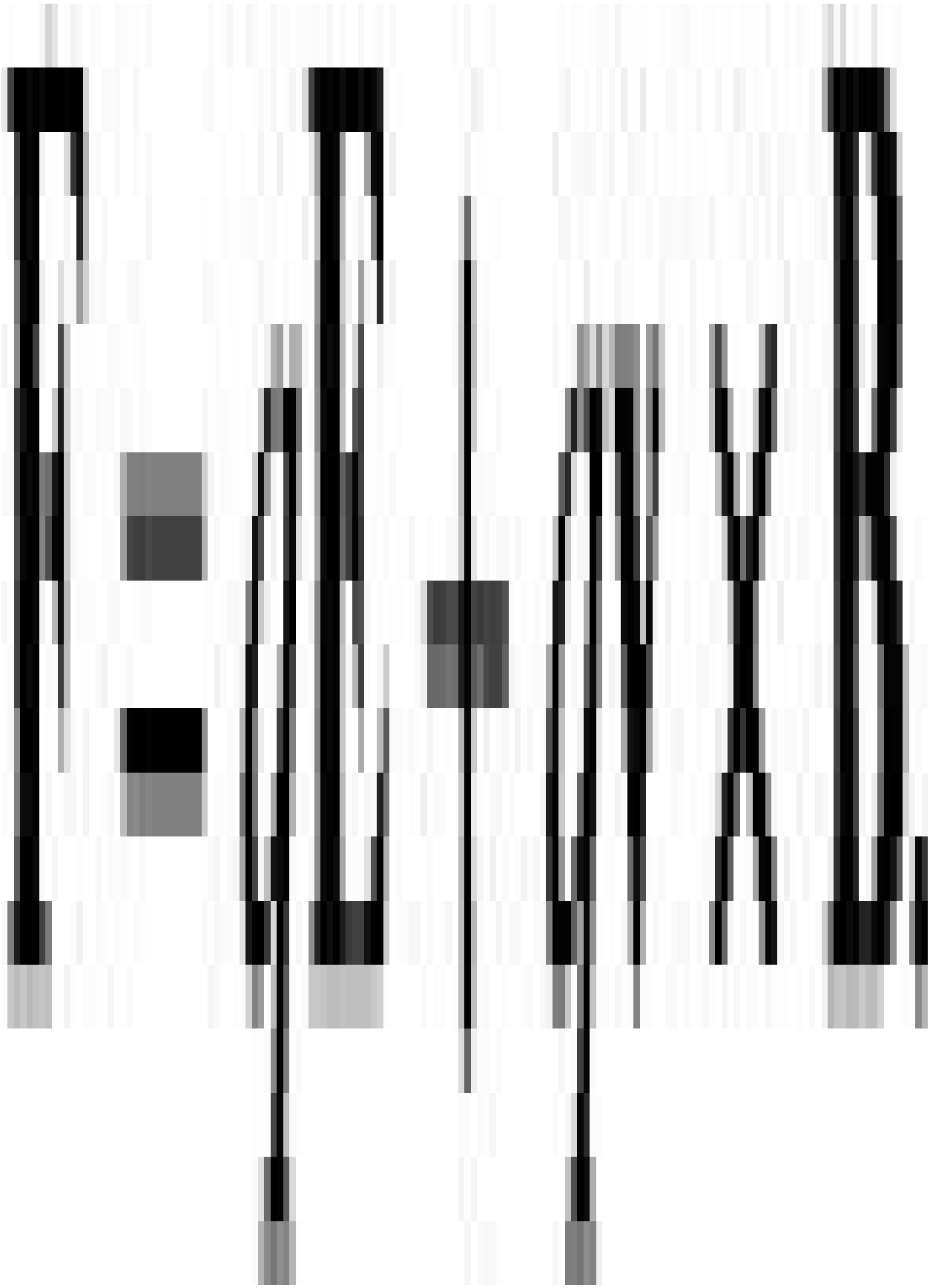
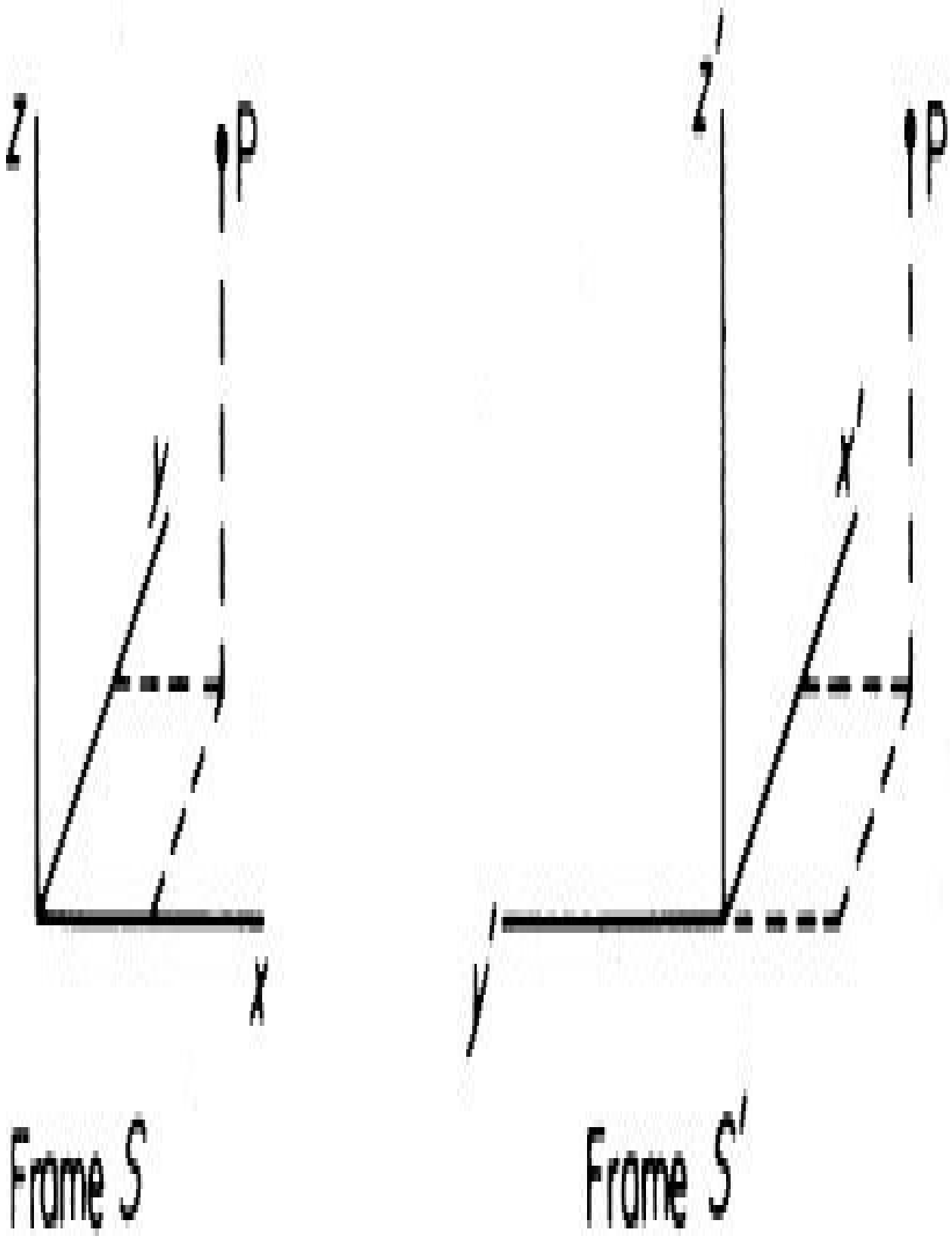


Figure 14.3. The frame S′ is generated by rotating frame S through 90° about the z-axis.

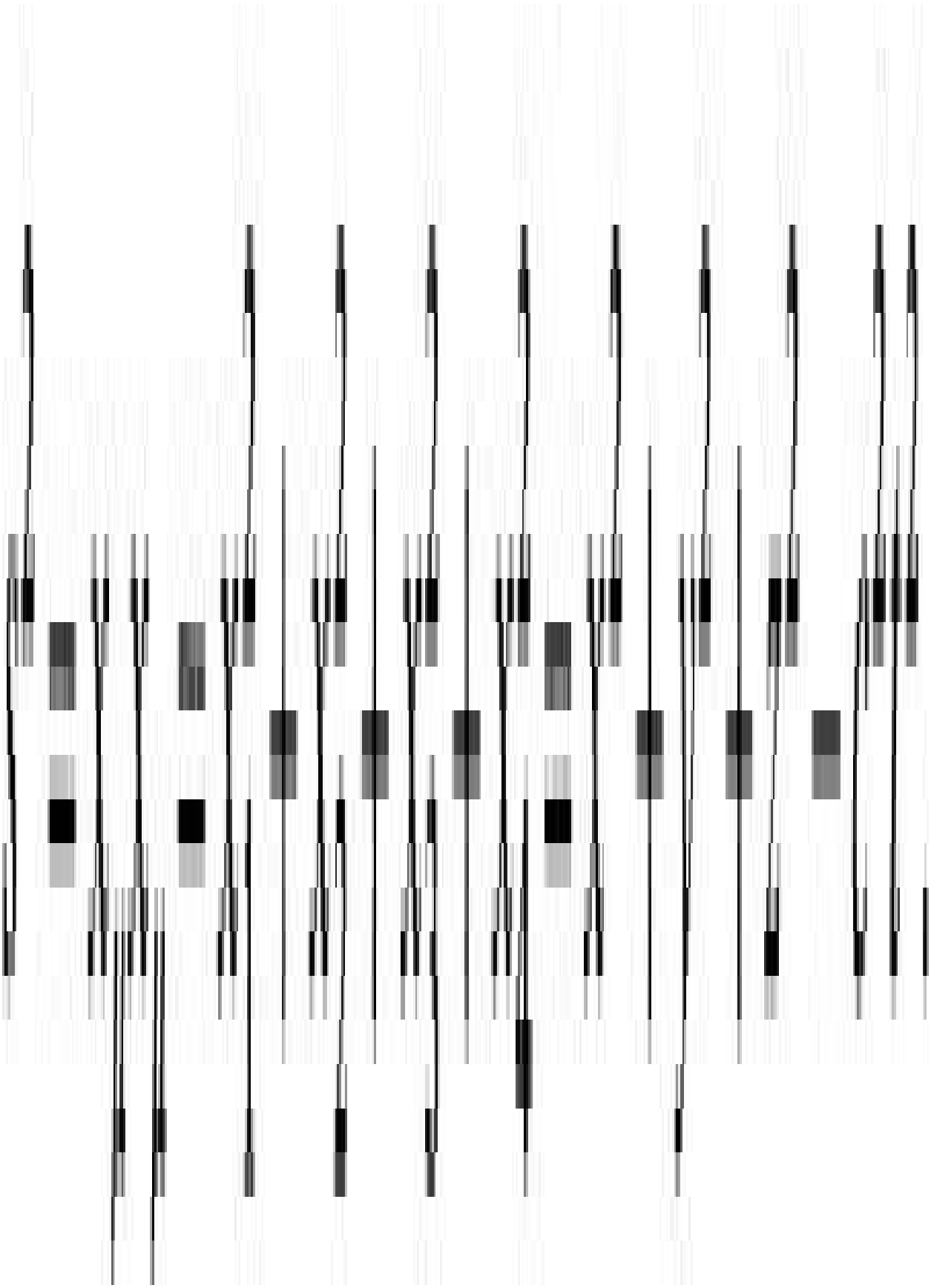

z
P
y
x
Frame S
z'
P
x'
y'
Frame S'

Each term on both sides of the equation is a vector, and the whole equation transforms in the same way as r. If the equation is true for one set of axes, it must also be true for axes at any other orientation. There is no need to bother working out components for a number of different axes: just by writing the equation in terms of vectors we have ensured that the result is independent of the orientation of the axes.

The fact that vector equations are not affected by rotating the coordinate system may seem obvious, but it is worth spelling out in detail because we now want to apply the same idea to the less easily visualized transformations of relativity. Like the rotational transformation, the Lorentz transformation is a linear relation between different coordinate systems, but it mixes up time with the spatial coordinates. An event occurring at a particular time and place in a frame of reference S is specified by the four coordinates (x,y,z) and t. The Lorentz transformation tells us the coordinates (x′, y′, z′) and t′ of the same event in a frame S′ moving relative to S. The event itself can be regarded as a four-dimensional vector, or four-vector for short.

The three dimensional vector r is characterized by its direction and length. The square of the length is found from the scalar product $\mathbf{r}\cdot\mathbf{r} = x^2 + y^2 + z^2$; $(x^2 + y^2 + z^2)$ has the same value for all orientations of the axes, i.e. it is invariant under rotations of the axes. A similar invariant scalar product can be formed from the components of a four-vector. In section 14.2 we have seen that the quantity $s^2 = x^2 + y^2 + z^2 - c^2t^2$ is Lorentz invariant, i.e. that s^2 has the same value when worked out in different inertial frames which are connected by a Lorentz transformation. By analogy with the three-dimensional scalar product r^2, we interpret s^2 as the square of the ‘length’ of the four-vector (x, y, z, t). Notice that s^2 is not always positive, because of the term $(-c^2t^2)$. To make s^2 look more like a familiar scalar product, it is convenient to replace the time coordinate by a new variable jct, which has the dimensions of a length. We shall write the four-vector representing an event as x_μ, with the suffix μ taking the values 1 to 4 for the four components x1 = x, x2 = y, x3 = z, x4 = jct. The invariant scalar product s^2 is given by

(14.6)



(The sum over μ has not been written explicitly. We use the convention that whenever a Greek suffix appears twice in a single term of a four-vector expression, this suffix is assumed to be summed from 1 to 4.)

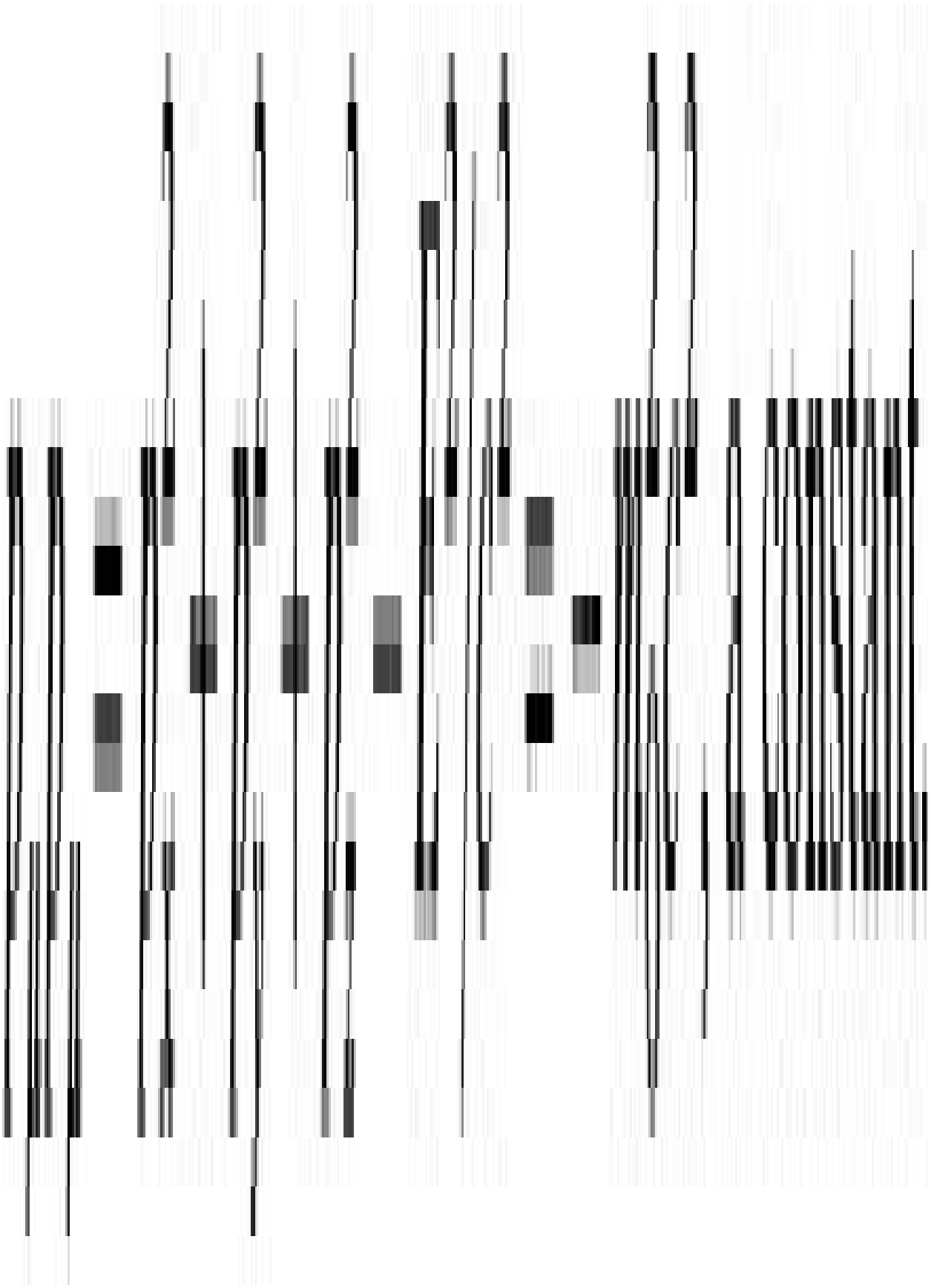
In the new notation, the Lorentz transformation is

(14.7)

and the inverse transformation, obtained by substituting (−v) for v, is

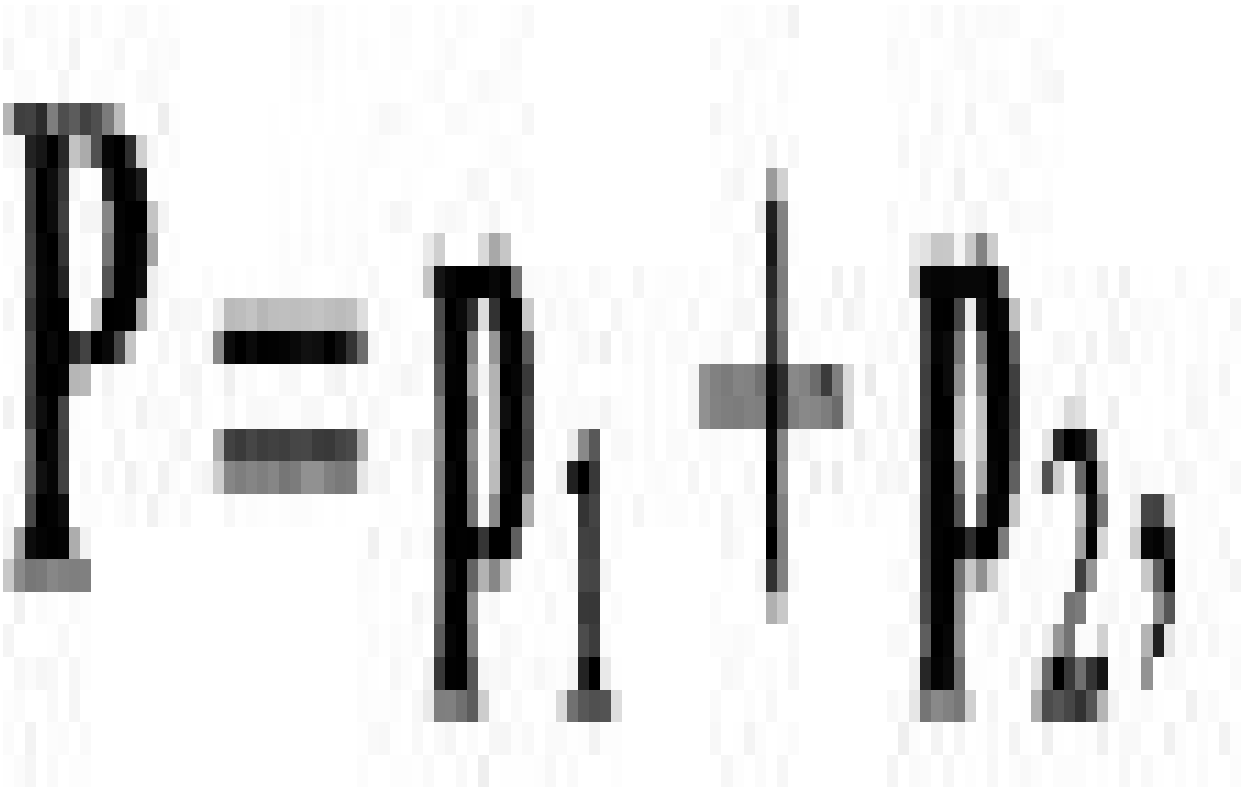
(14.8)

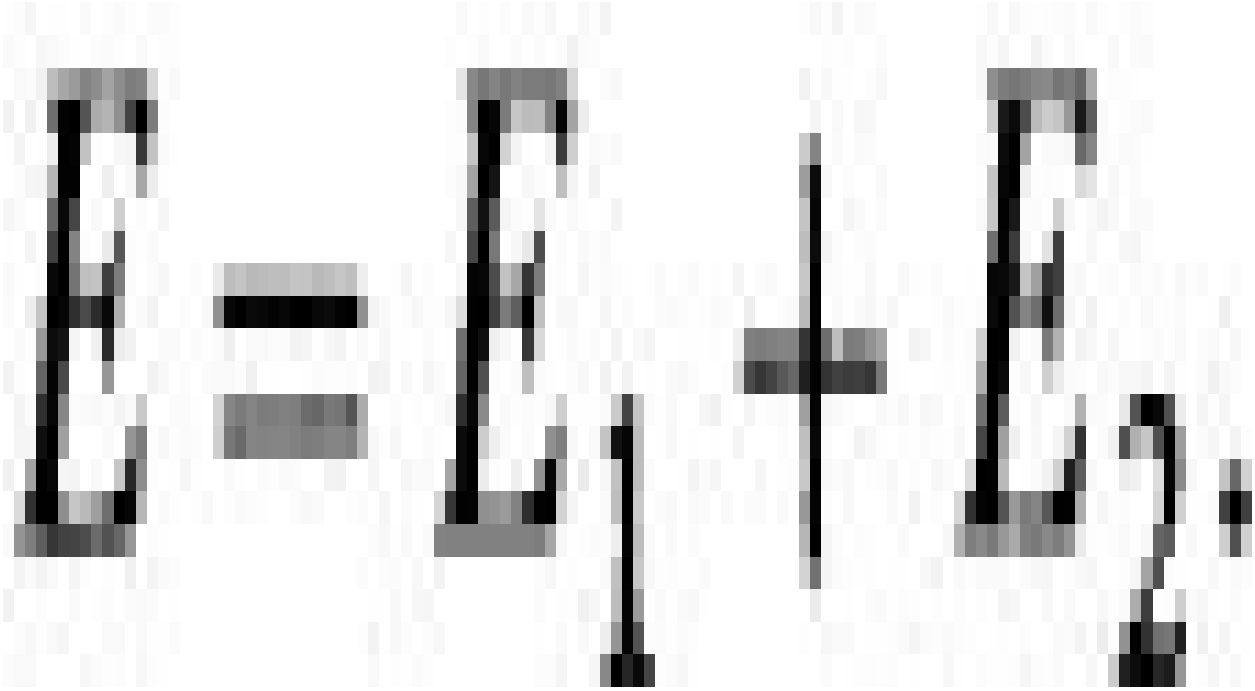
Just as there are other three-dimensional vectors besides position vectors, so there are other four-vectors besides those whose space and time coordinates represent an event. For example, the momentum and energy of a particle together form a four-vector p_μ with components (px, py, pz, jE/c). These components transform according to the Lorentz transformation, with the same coefficients as occur in Equation (14.7) for the four-vector x_μ. The scalar product $p_\mu p_\mu$ is invariant, i.e.



All observers in inertial frames agree about the value of the rest mass m0 of the particle, although they may differ about its momentum and energy.

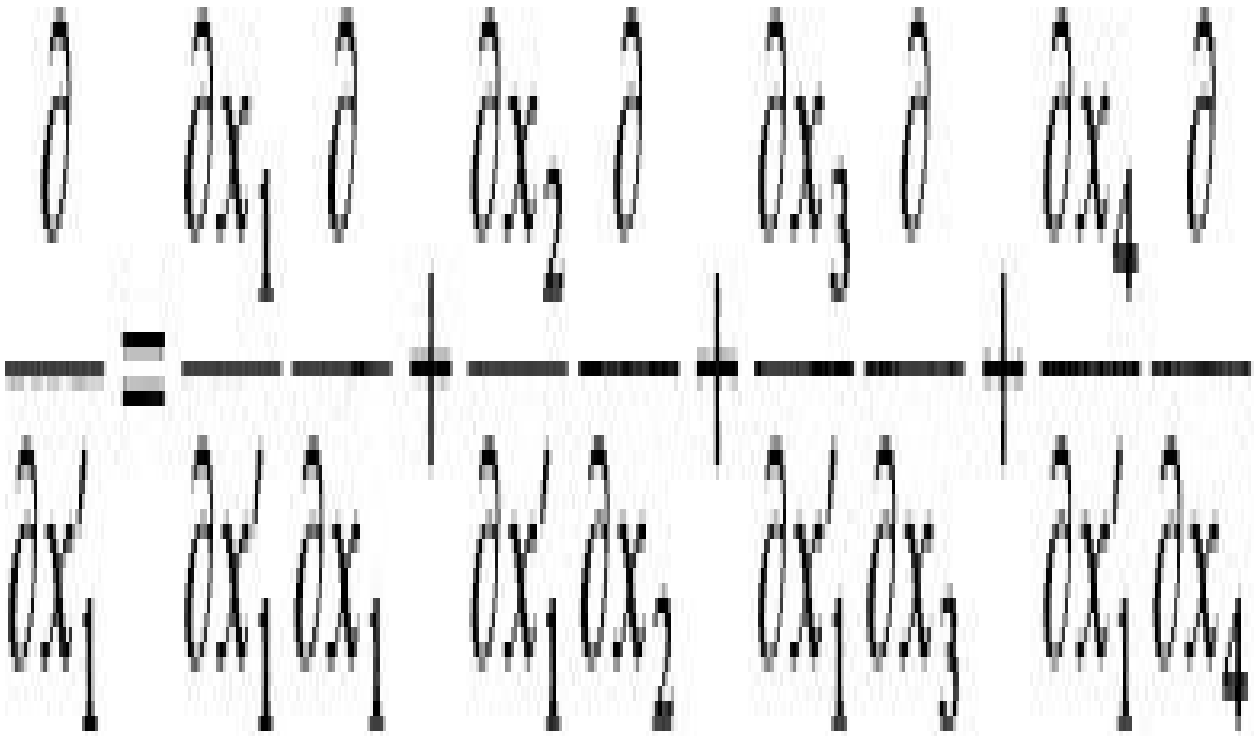
Four-vectors may be combined together to form four-vector equations, analogous to the vector equations in three dimensions. A trivial example of a four-vector equation is found in the addition of the momenta and energies of two particles. The total momentum P of the pair of particles is the sum of their individual momenta p1 and p2, and similarly the total energy E is the sum of their energies E1 and E2:

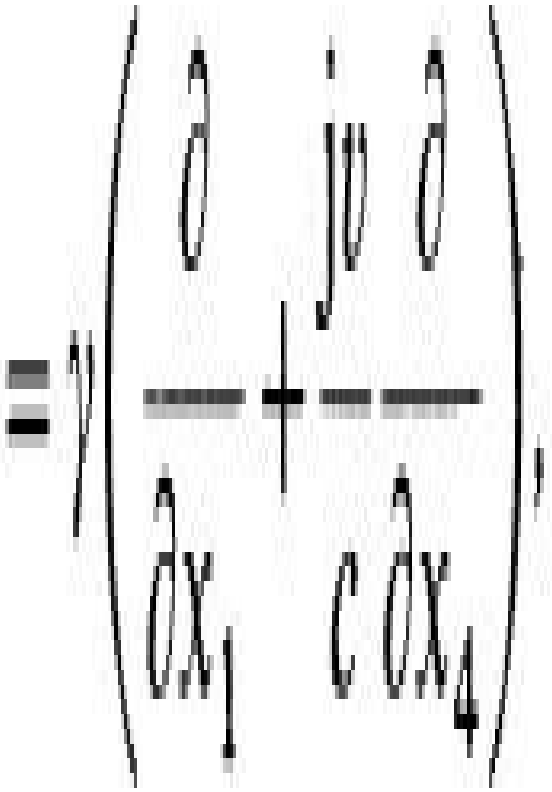




Because the Lorentz transformation is linear, (P, E) transforms in the same way as (P1, E1) and (p2, E2), or in other words P and E form the components of a new four-vector $P\mu$ defined by the four-vector equation $P\mu = p1\mu + p2\mu$.

In order to investigate the relativistic behaviour of Maxwell's equations, we shall need to know the transformation properties of differential operators. The space and time differential operators fit easily into the four-vector scheme, since together they form yet another four-vector. To see that this is so, we apply the rules of partial differentiation to write





using Equations (14.8). This is the same transformation law as applies to the coordinate itself. Similar equations for the other components show that they too transform according to the Lorentz transformation. In other words, $\partial/\partial x_\mu$ is a four-vector.

From the four vector $\partial/\partial x_\mu$ we can form the invariant scalar product

$$\frac{\partial}{\partial x_\mu}\frac{\partial}{\partial x_\mu} = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial x_3^2} + \frac{\partial^2}{\partial x_4^2}$$

$$= \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} - \frac{1}{c^2}\frac{\partial^2}{\partial t^2},$$

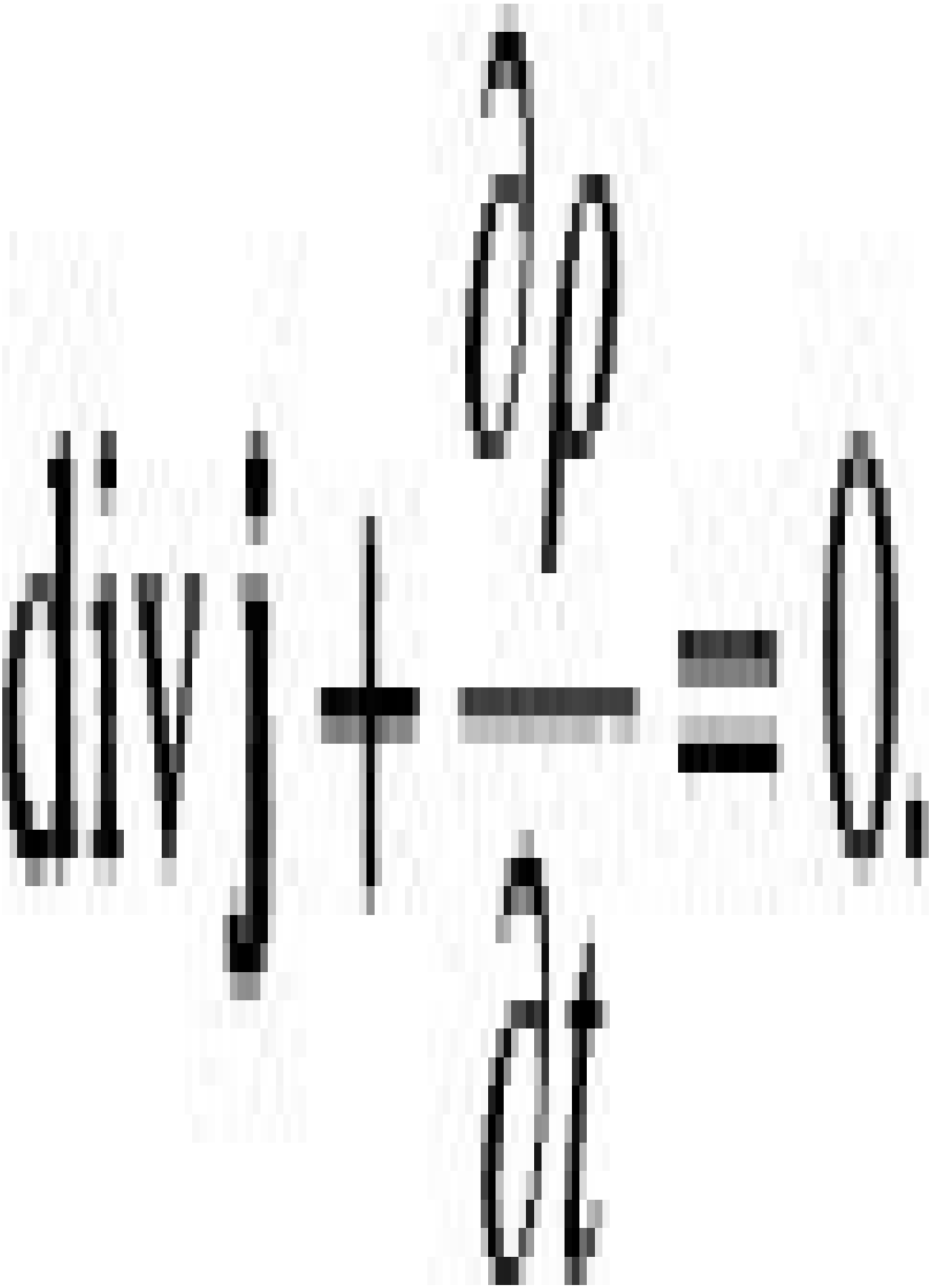
This is the operator occurring in the wave equation for waves moving at a speed c: it must be an invariant, since all observers agree about the speed of light.

14.5 MAXWELL'S EQUATIONS IN FOUR-VECTOR FORM

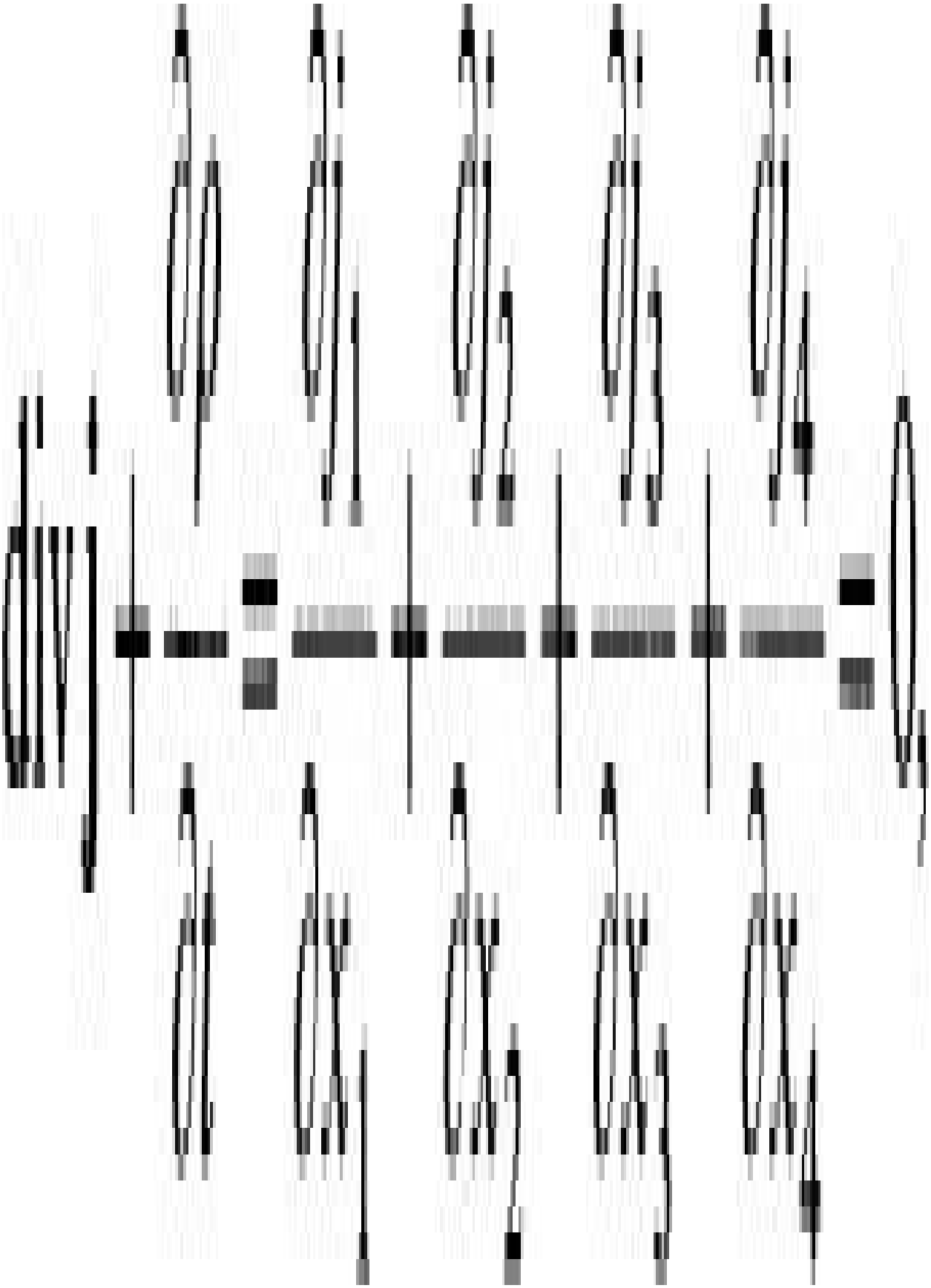
In section 14.3 we found that electric and magnetic fields transform into one another in different inertial frames. Electric and magnetic fields are therefore not simply the space-like components of two different four-vectors. In the four-dimensional picture the behaviour of the fields is rather complicated—together they form some of the components of a quantity called a four-tensor. For our purposes there is fortunately no need to know anything about tensors, since it turns out that the vector and scalar potentials do make up a four-vector. By expressing Maxwell's equations in terms of the potentials, we shall be able to write them in a way which is obviously relativistically invariant. But before doing this, we must first show that current density and charge density also constitute a four-vector, and write the equation of continuity in an invariant form.

The equation of continuity

We have already seen that the charge on any body is invariant, so that it is assigned the same value whatever may be the velocity of the body relative to an observer measuring the charge. Charge is also conserved, since a charged elementary particle can only be created or annihilated in association with another particle carrying charge of exactly the same magnitude but opposite size: this must evidently be true in all inertial frames. Now the mathematical statement of charge conservation is contained in the equation of continuity (Equation (10.1)) and this equation must therefore have the same form in all inertial frames. In the frame S, the equation of continuity is



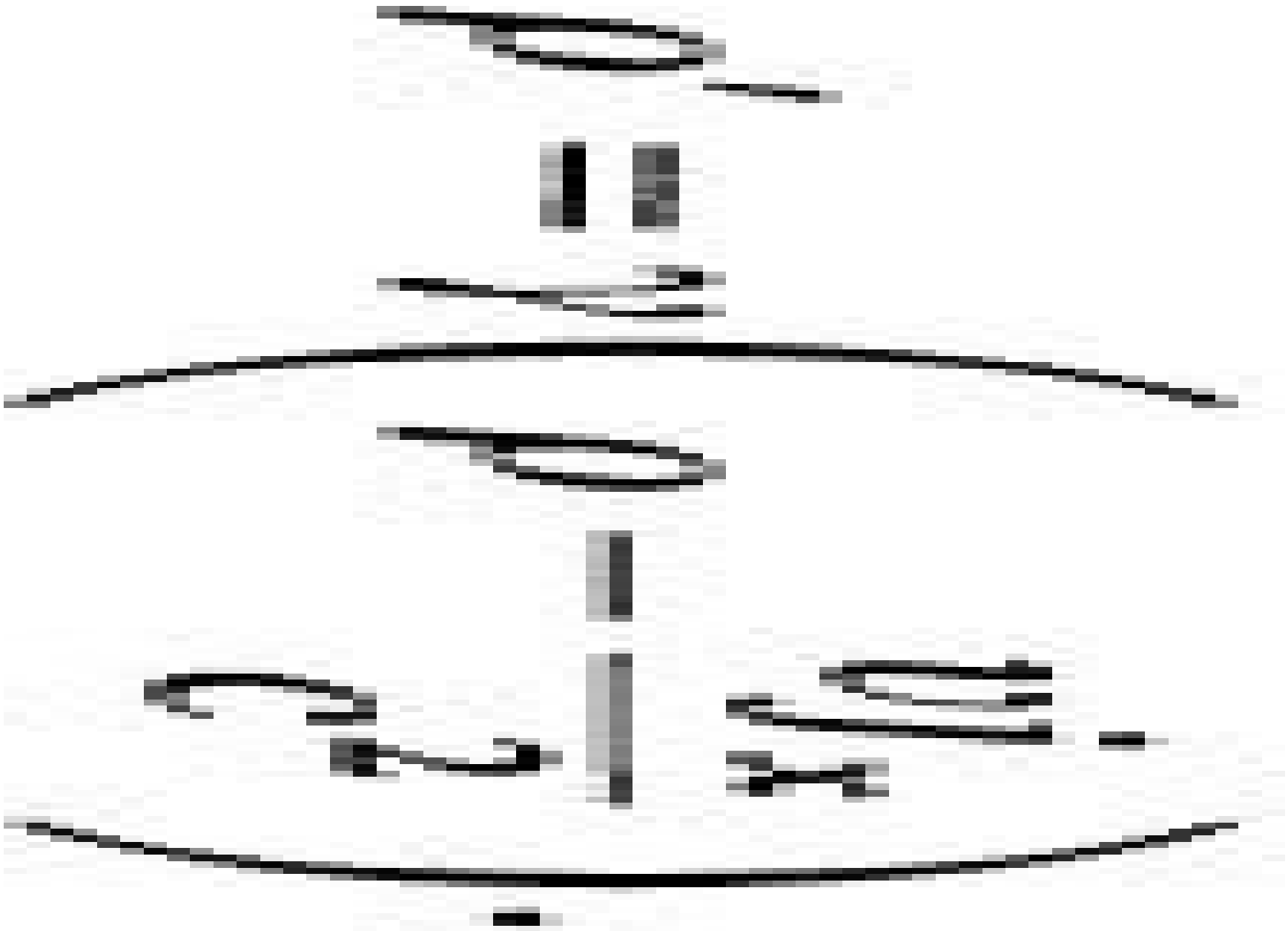
This equation takes on the required invariant form only if we introduce a four-vector $j\mu$ made up from the current density and the charge density. The current density j is the space-like part of $j\mu$, and the fourth component is $j4 = jc\rho$. Rewriting the equation of continuity in terms of $j\mu$,



or

(14.9)

The equation of continuity is now in four-vector form, and manifestly invariant. As we are insisting from the start that the law of charge conservation must be invariant, Equation (14.9) constitutes a proof that (j, jcρ) is indeed a four-vector. Its components transform in the same way as the components of (r, jct). Thus in the frame S′, moving at a speed v along the x-axis of S,



In the example of the line charges discussed in section 14.3, the charges were at rest in frame S, and $j_x = 0$. As a result of the FitzGerald contraction, the transformation is then simply $\rho' = \gamma\rho$.

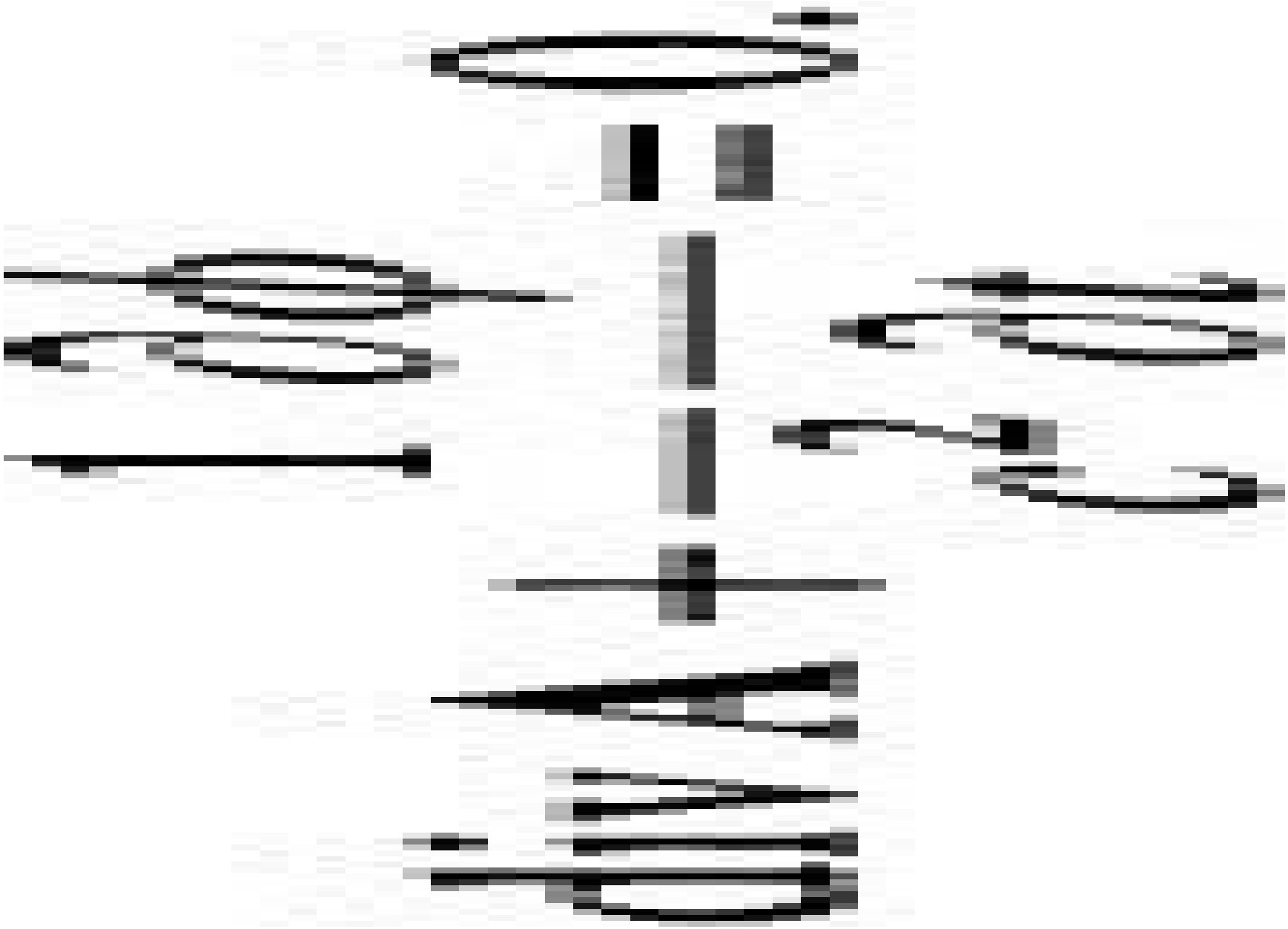
Maxwell's equations

To show that Maxwell's equations are Lorentz invariant, it is sufficient to express them in four-vector form. We have already pointed out that the electromagnetic fields are not components of four-vectors, and we can most easily test the invariance of Maxwell's equations by writing them in terms of the scalar and vector potentials. This has already been done in the previous chapter for the case when no dielectric or magnetic materials are present; we shall continue to make this restriction in order to avoid undue complication. The fields are derived from the potentials using Equations (13.1) and (13.2):

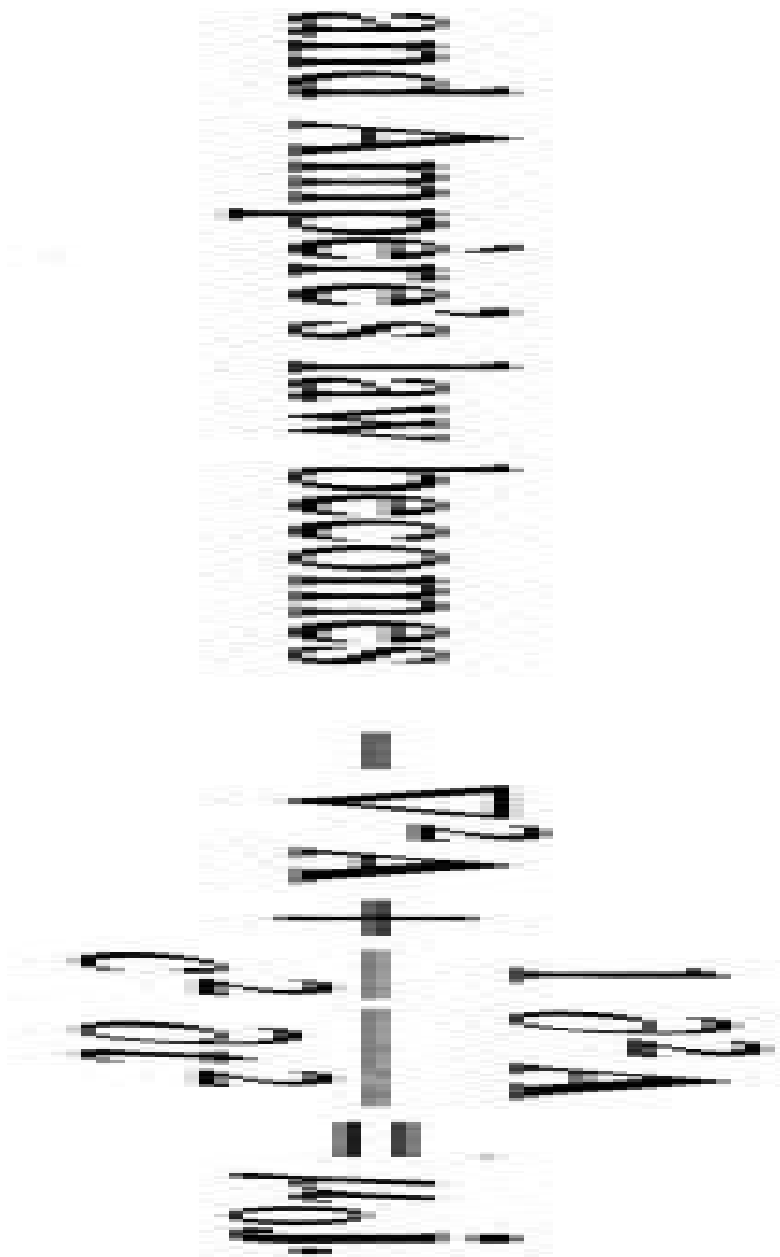
$$\mathbf{B} = \operatorname{curl} \mathbf{A}$$

$$\mathbf{E} = -\operatorname{grad} \phi - \frac{\partial \mathbf{A}}{\partial t}.$$

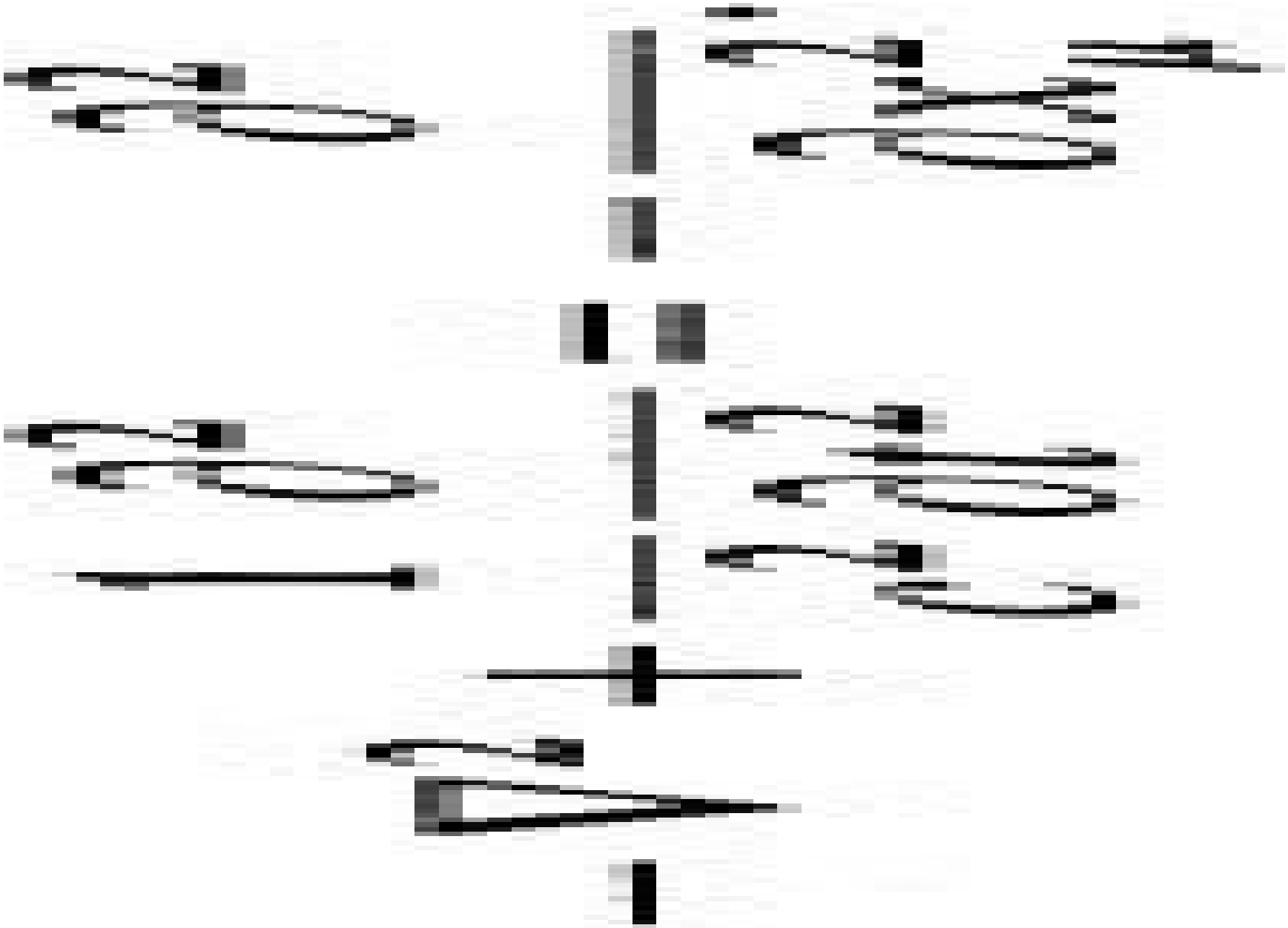
Because of the vector identities div curl A ≡ 0 and curl grad $\phi \equiv 0$, the law that no free magnetic poles exist (div B = 0) and Faraday's law (curl $E = -\partial B/\partial t$) are automatically satisfied. However, there is still some arbitrariness in the potentials, which can have a variety of different forms while still remaining consistent with Maxwell's equations. Most of the possible choices of potentials will not lead to a Lorentz invariant form for the equations describing how A and ϕ vary with r and t. To achieve invariance, we must require the potentials to satisfy the Lorentz condition (Equation (13.8))



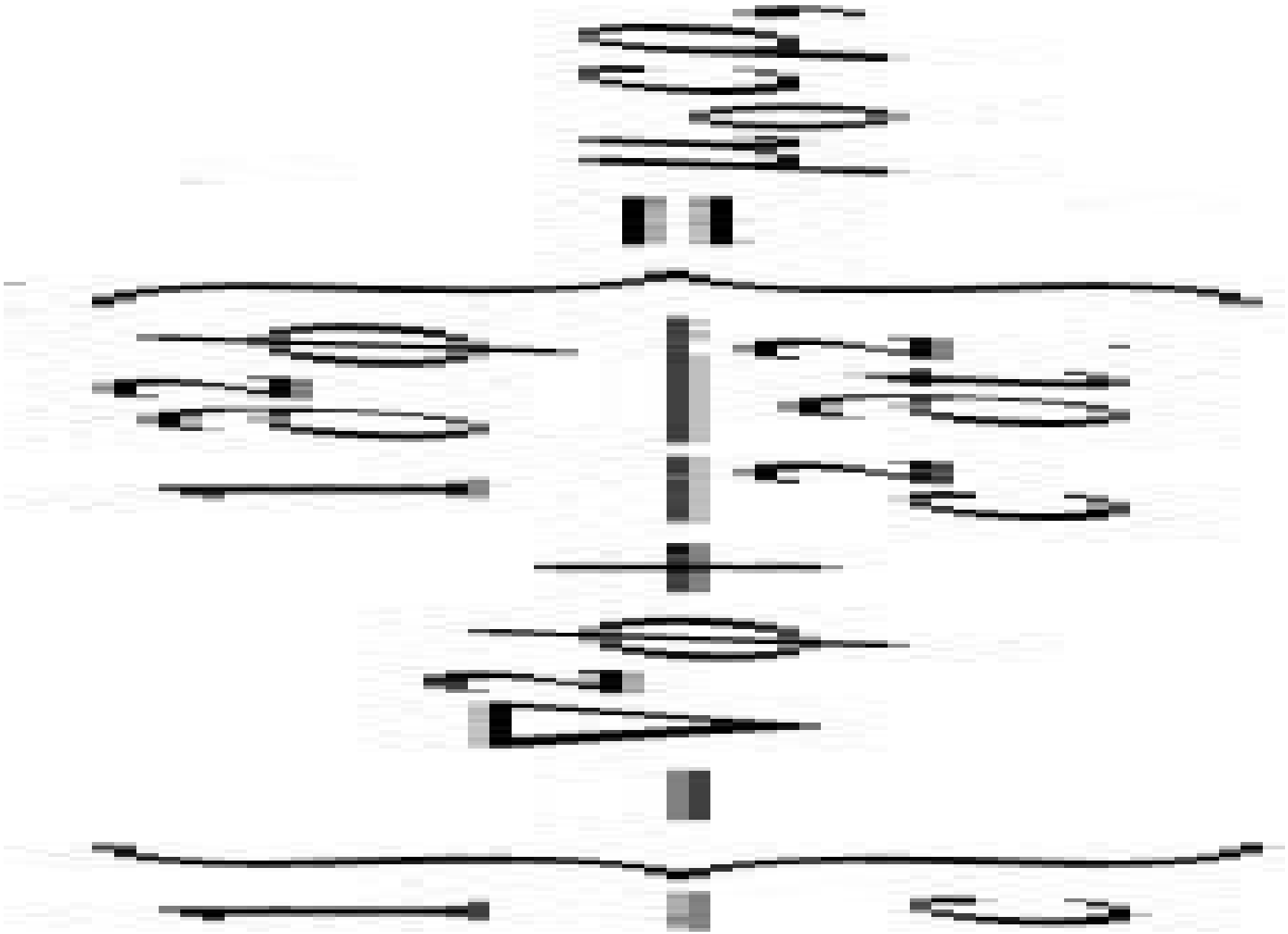
This restriction on the potentials was chosen in section 13.1, where it was shown to lead to the following forms for the remaining two of Maxwell's equations:



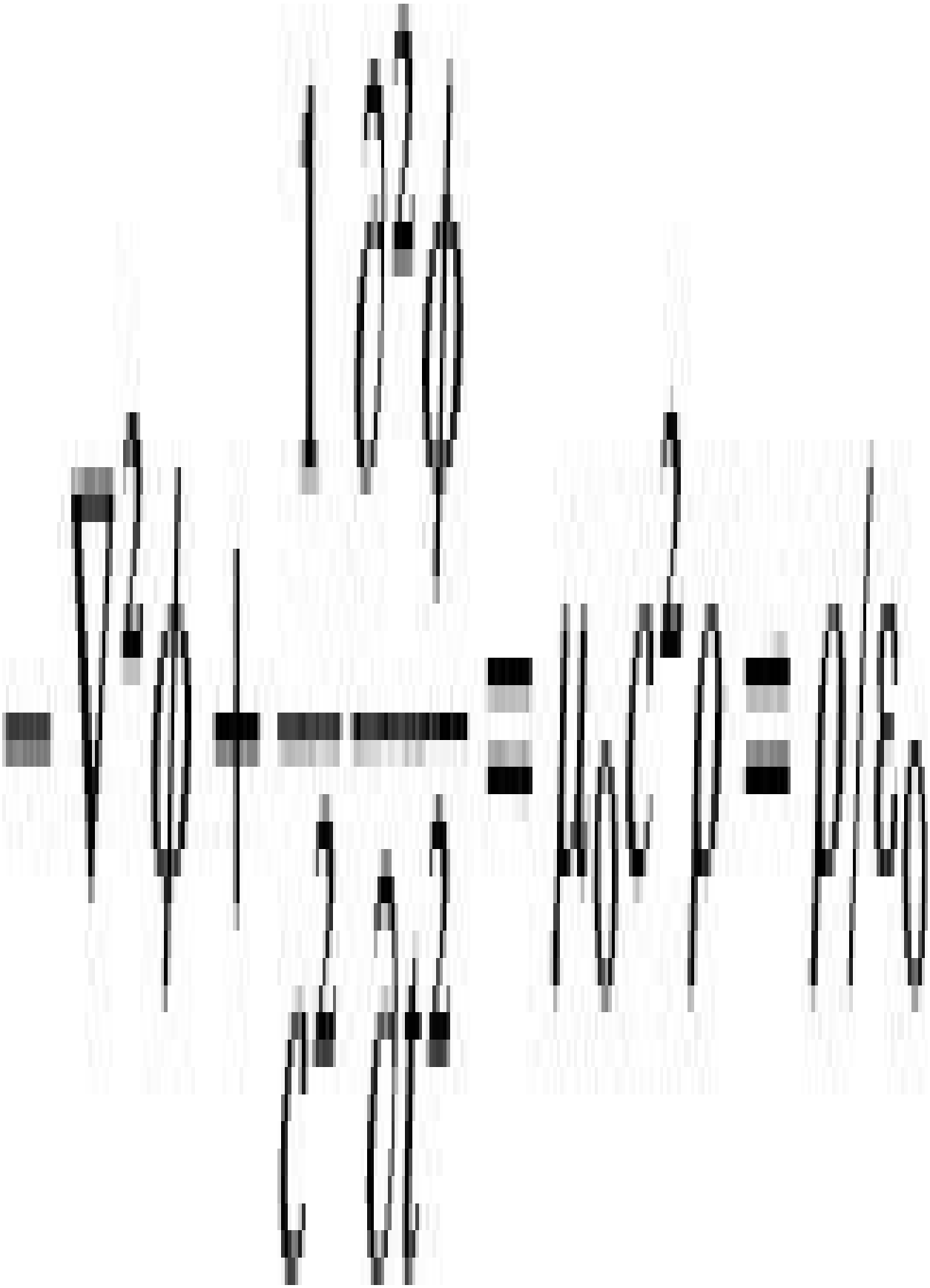
Apart from constant factors, the right-hand sides of these equations are the components of the four-vector $jv = (j, jc\rho)$. On the left-hand sides the operator



operates on the scalar and vector potentials. This operator is Lorentz invariant, and under Lorentz transformations both sides of the equations transform in the same way only if the quantity $Av = (A, j\phi/c)$ is a four-vector. By comparison with Ampère's law, the fourth component equation is found by making the substitution $A \rightarrow j\phi/c$, $j \rightarrow jc\rho$, i.e.



or



which is indeed Gauss' law. Ampère's law and Gauss' law can then be written together in the single four-vector equation

(14.10)

Notice that the Lorentz condition itself is also invariant. It can now be written in terms of four-vectors, and is

(14.11)

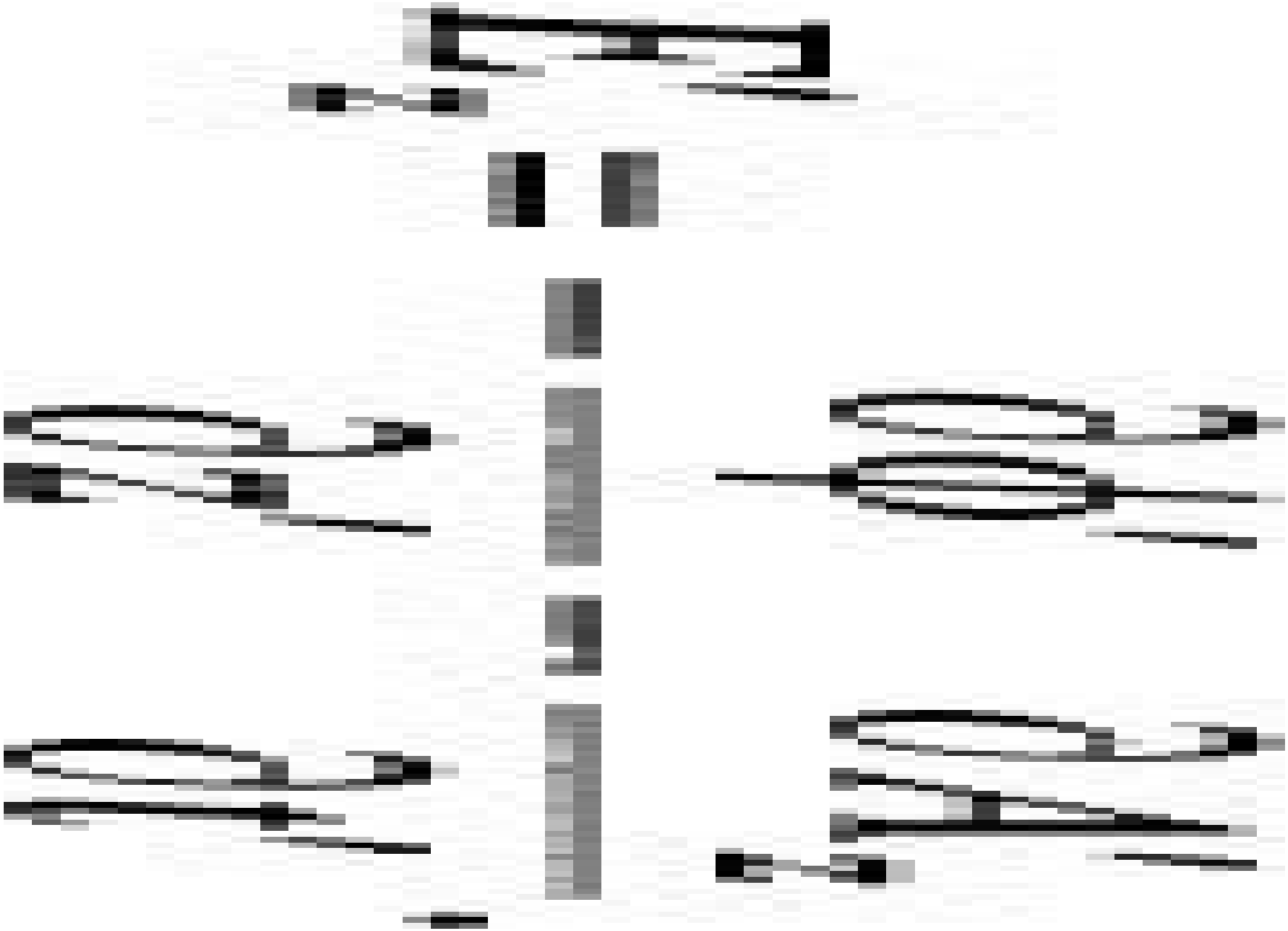
To sum up, we have shown that without any modifications at all, Maxwell's equations can be written in Lorentz invariant form, as required by the first axiom of the special theory of relativity. Maxwell's equations are also automatically in accordance with the second axiom that all observers measure the speed of light to be c, since in free space Equation (14.10) reduces to a wave equation for electromagnetic waves travelling at the speed c. As well as finding that Maxwell's equations are consistent with the special theory of relativity, we have also learnt something new, namely that the quantities Av and jv are four-vectors. Knowing Av and jv in one inertial frame, we can now work out their values in any other inertial frame by using the Lorentz transformation.

We have not yet completed the test of electromagnetism against relativity, because the equation defining the Lorentz force $F = q(E + v \times B)$ has not so far been shown to be invariant. Since the Lorentz force is expressed in terms of the fields E and B, we must first investigate the transformation of the fields, and then make sure that the force itself transforms in the proper manner. This is done in the next section.

14.6 TRANSFORMATION OF THE FIELDS

Electric and magnetic fields do not form the space-like parts of four-vectors. However, the fields can be derived by differentiating the vector and scalar potentials. Knowing how to transform the four-potential $A\mu$ and the differential operator $\partial/\partial\mu$, we can work out the transformation of the fields one component at a time. For example, let us evaluate the electric field component $E'z$ in the S' frame in terms of the fields E and B in the S frame.

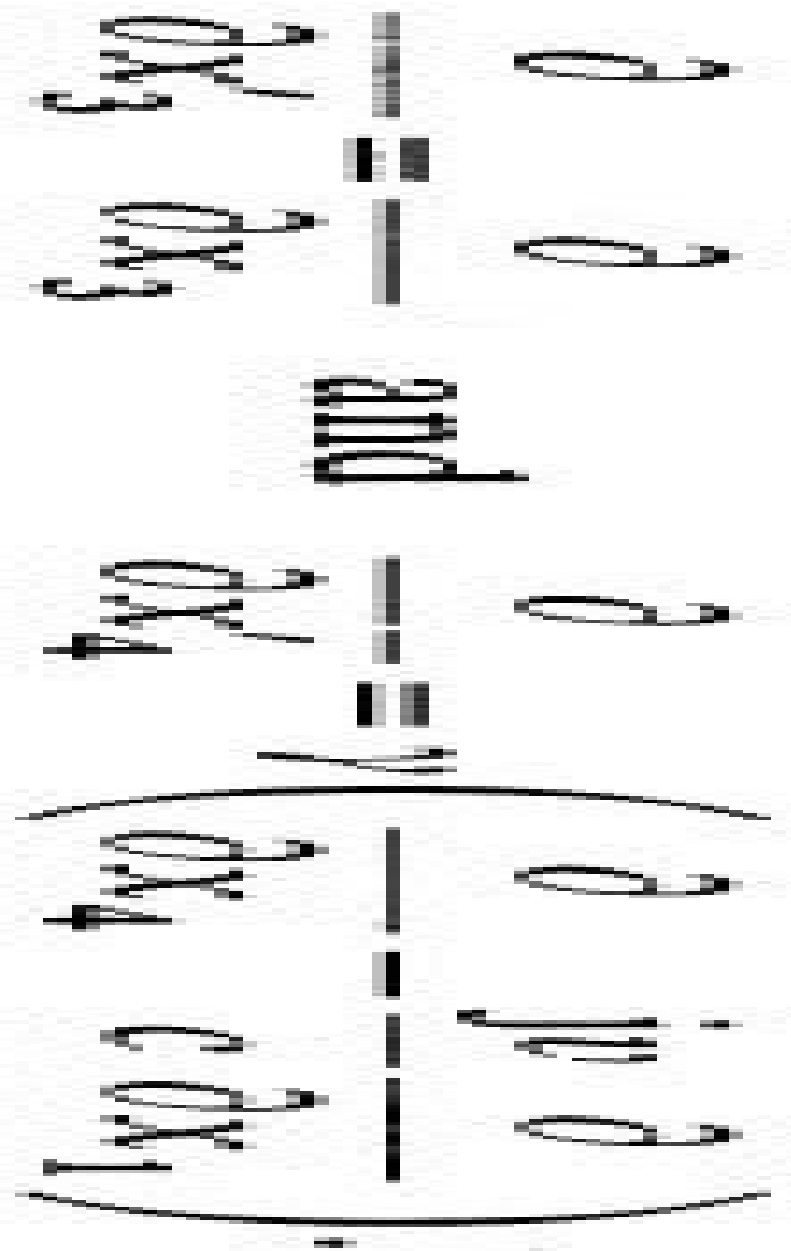
Now



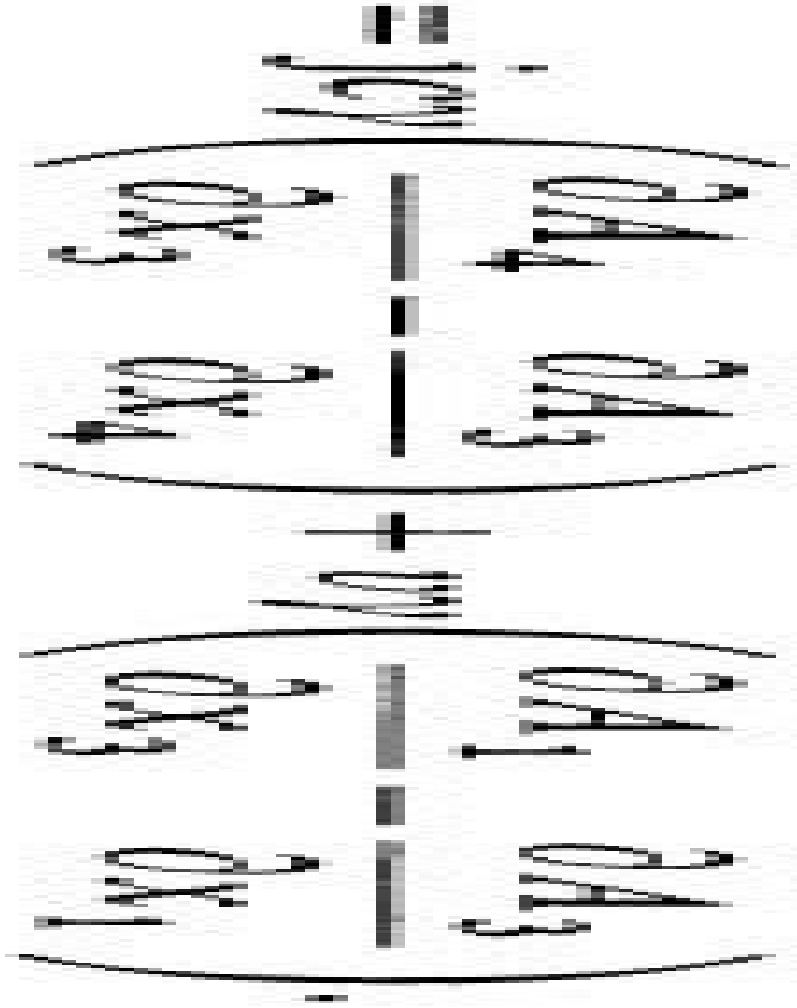
Since $A\mu = (A, j\phi/c)$ and $\partial/\partial x\mu = (\nabla, (j/c)\partial/\partial t)$ are both four-vectors, this equation may be rewritten as

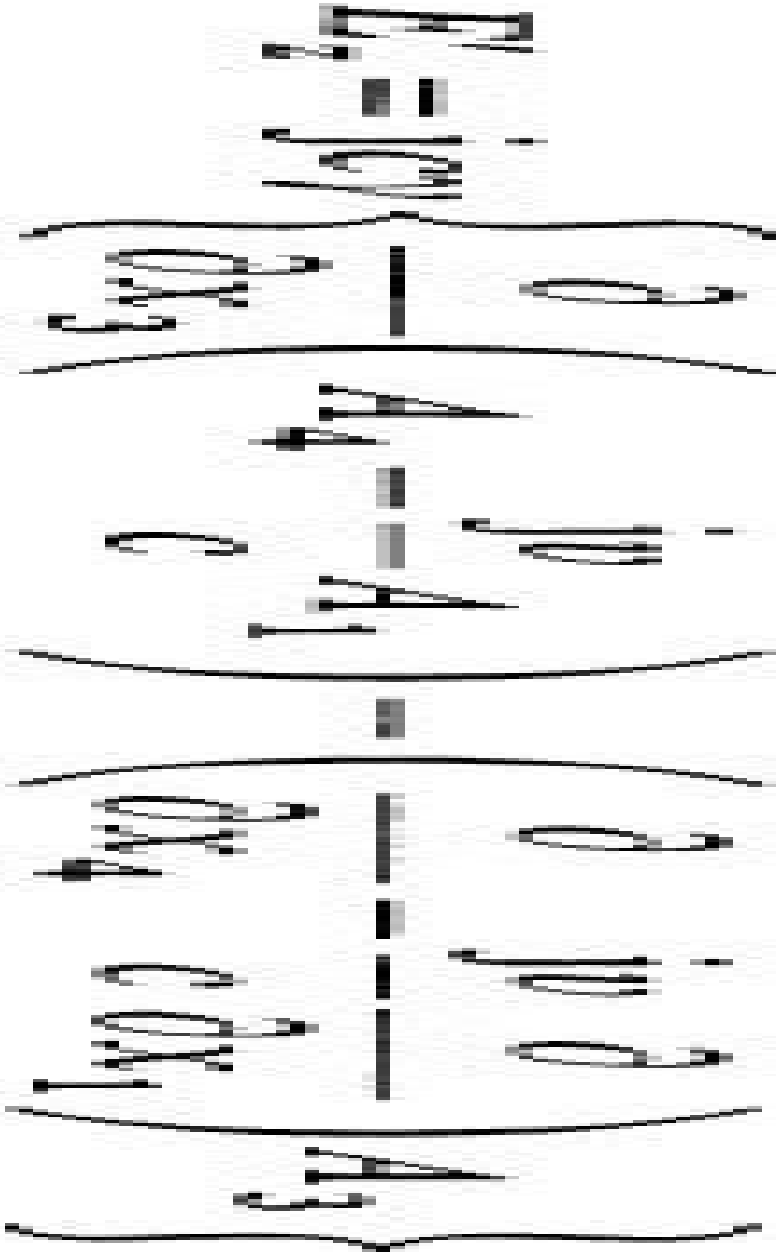
(14.12)

Using the Lorentz transformation given in Equation (14.7),

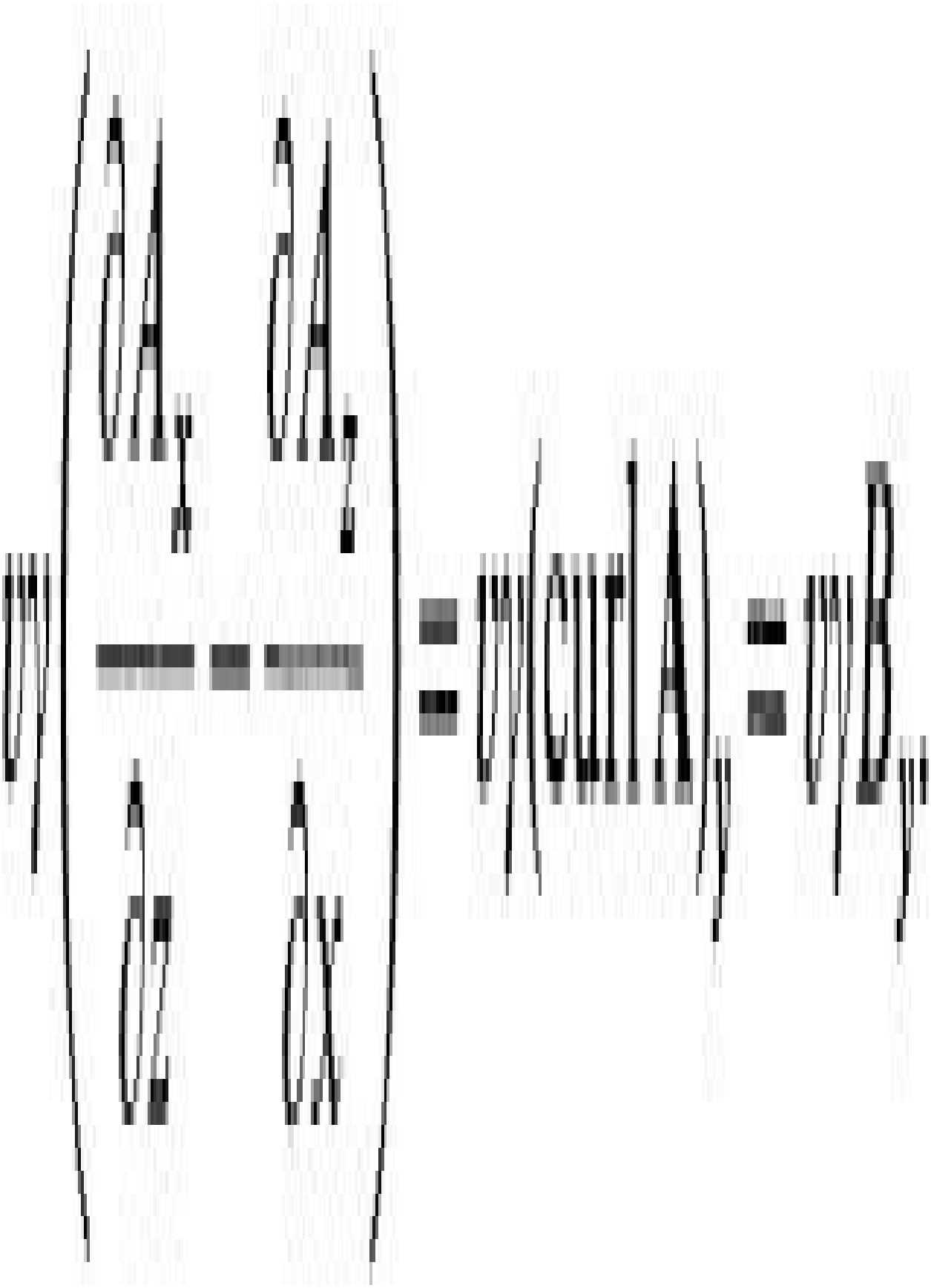


Hence





By comparison with Equation (14.12), the first term is seen to be simply γEz. The second term looks unfamiliar in its four-vector disguise, but it contains only space-like quantities, and is



Finally we have $E'_z = \gamma(E_z + vB_y)$.

This is in agreement with the result derived earlier (see Equation (14.5)) for the special case in which there is no electric field in the S frame.

Proceeding in the same way, all the other components of E' and B' can be evaluated in terms of the components in the S frame. The results are

(14.13)

and

(14.14)

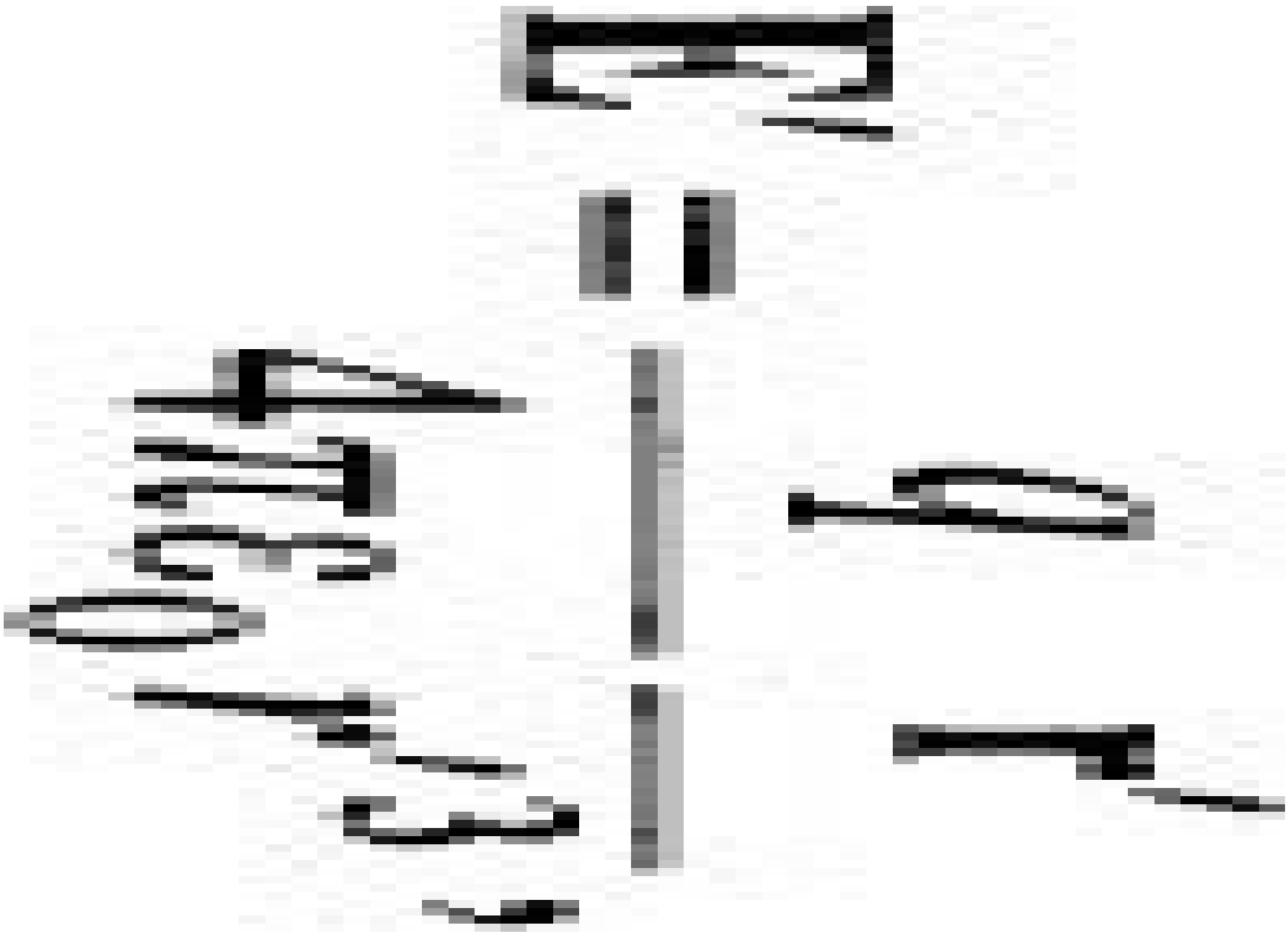
The inverse transformations for E and B in terms of E' and B' have the same form except that v is replaced by $(-v)$.

For a macroscopic body moving at a speed small compared with c, there is only an insignificant difference between the electric and magnetic fields as measured by a stationary observer and by one moving with the body. Suppose, for example, that an object is travelling at the speed of sound (300/m sec) perpendicularly to an electric field of magnitude 10^6 volts/m—about as large a field as can be maintained over a distance of a few metres in vacuo. According to Equation (14.14), an observer moving with the object sees not only an electric field, but also a magnetic field of magnitude $\gamma vE/c^2$ 3×10^{-9}T. This is

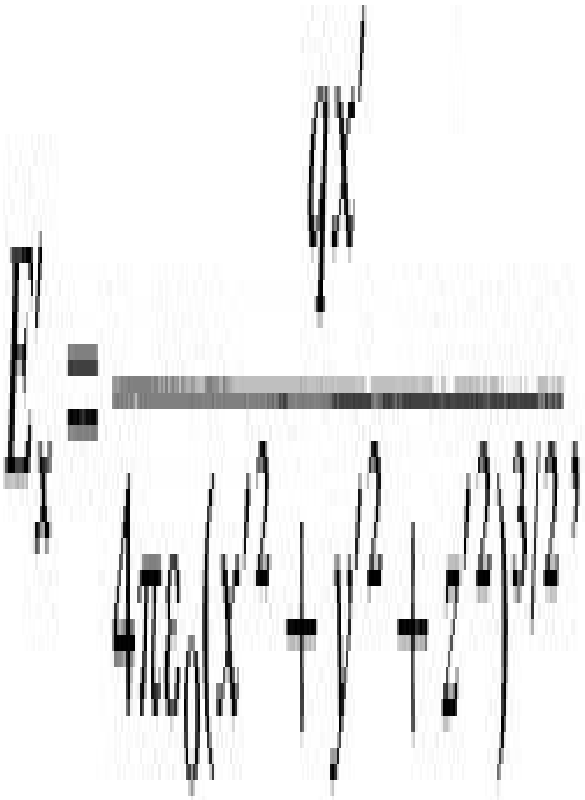
completely negligible even by comparison with the earth's rather weak field. On the atomic scale, on the other hand, speeds near to c are attained, and the role of external electric and magnetic fields can be very much modified. Consider a neutral hydrogen atom with an energy of 100 keV in the laboratory frame; a neutral atom can be given this energy by accelerating a proton through 100 kV and subsequently allowing it to capture an electron. The speed of such an atom is roughly 5×10^6 m/sec. If the atom now enters a magnetic field of 1 T, then in the frame S′ at rest with respect to the atom, Equation (14.13) tells us that there is a huge electric field of magnitude 5×10^6 volts/m! The neutral atom, which has only recently captured its electron, is most likely to be in an excited state, and when subjected to such a strong electric field, it will very quickly ionize again*. This process is called Lorentz ionization.

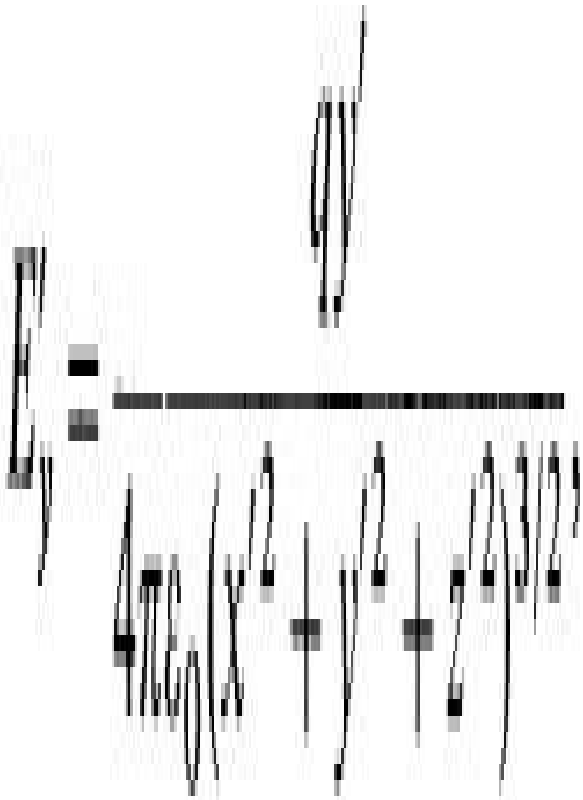
Fields due to a moving point charge

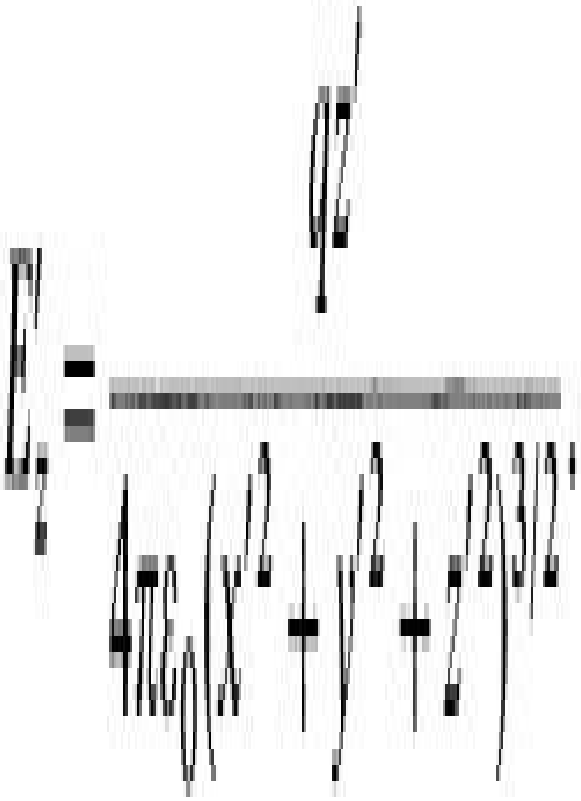
The field transformations allow us to work out the fields generated by a point charge q moving with a constant velocity. Choosing the x-axis to be in the direction of motion of the charge, we can arrange that the charge is at rest at the origin of the S′ frame. In this frame there is no magnetic field, and the electric field is simply



with components







The components of the fields in S now follow from the inverse of the transformations given in Equations (14.13) and (14.14) together with the Lorentz transformation as in Equation (14.1). We have

(14.15)

$$E_x = E'_x = \frac{qx'}{4\pi\varepsilon_0(x'^2 + y'^2 + z'^2)^{3/2}}$$

$$= \frac{q\gamma(x - vt)}{4\pi\varepsilon_0[\gamma^2(x - vt)^2 + y^2 + z^2]^{3/2}},$$

$$E_y = \gamma E'_y = \frac{q\gamma y}{4\pi\varepsilon_0[\gamma^2(x - vt)^2 + y^2 + z^2]^{3/2}},$$

$$E_z = \gamma E'_z = \frac{q\gamma z}{4\pi\varepsilon_0[\gamma^2(x - vt)^2 + y^2 + z^2]^{3/2}},$$

and

$$B_x = 0,$$

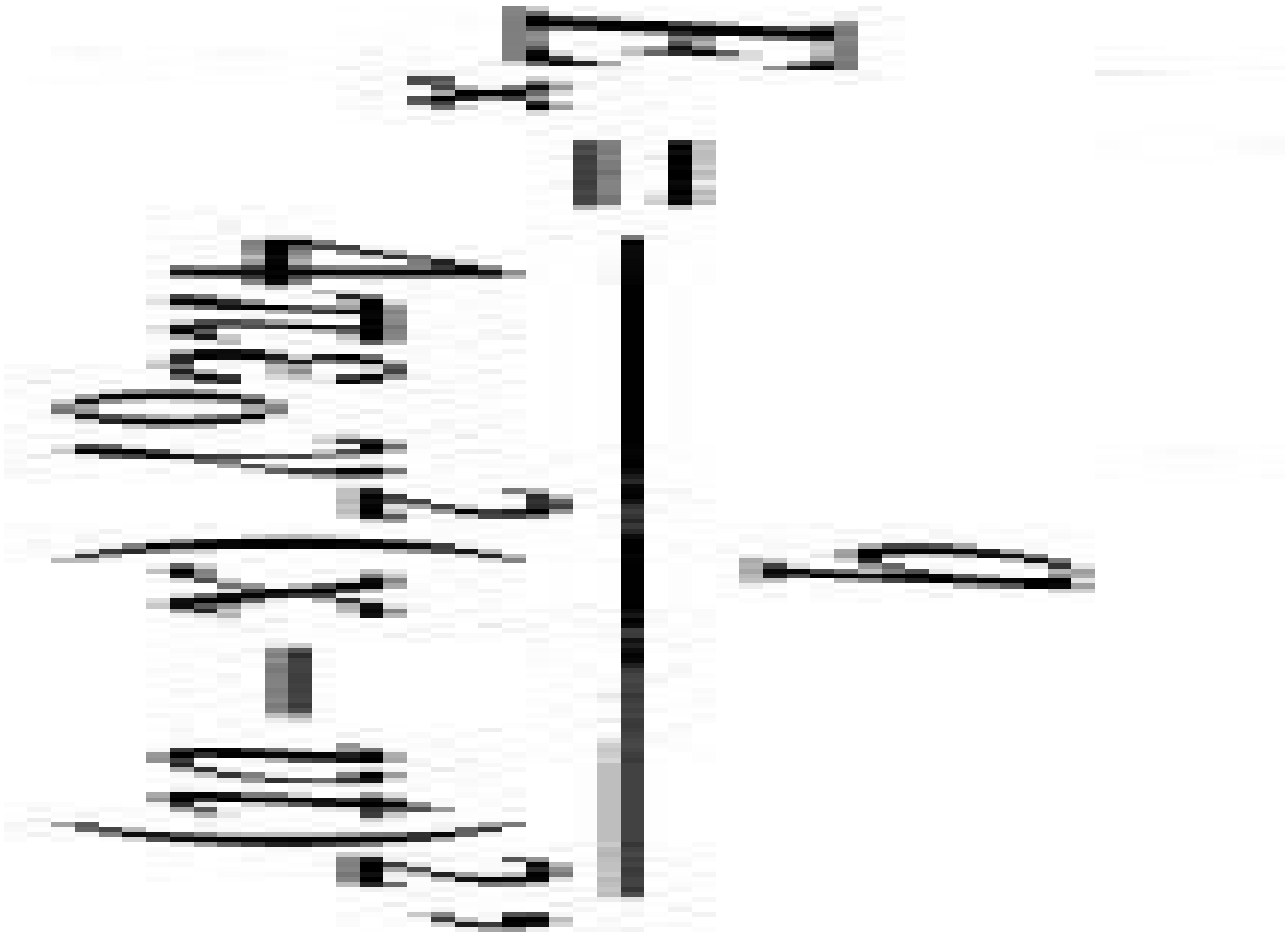
$$B_y = -\frac{\gamma v E'_z}{c^2} = -\frac{v E_z}{c^2}$$

$$B_z = \frac{\gamma v E'_y}{c^2} = \frac{v E_y}{c^2},$$

or equivalently

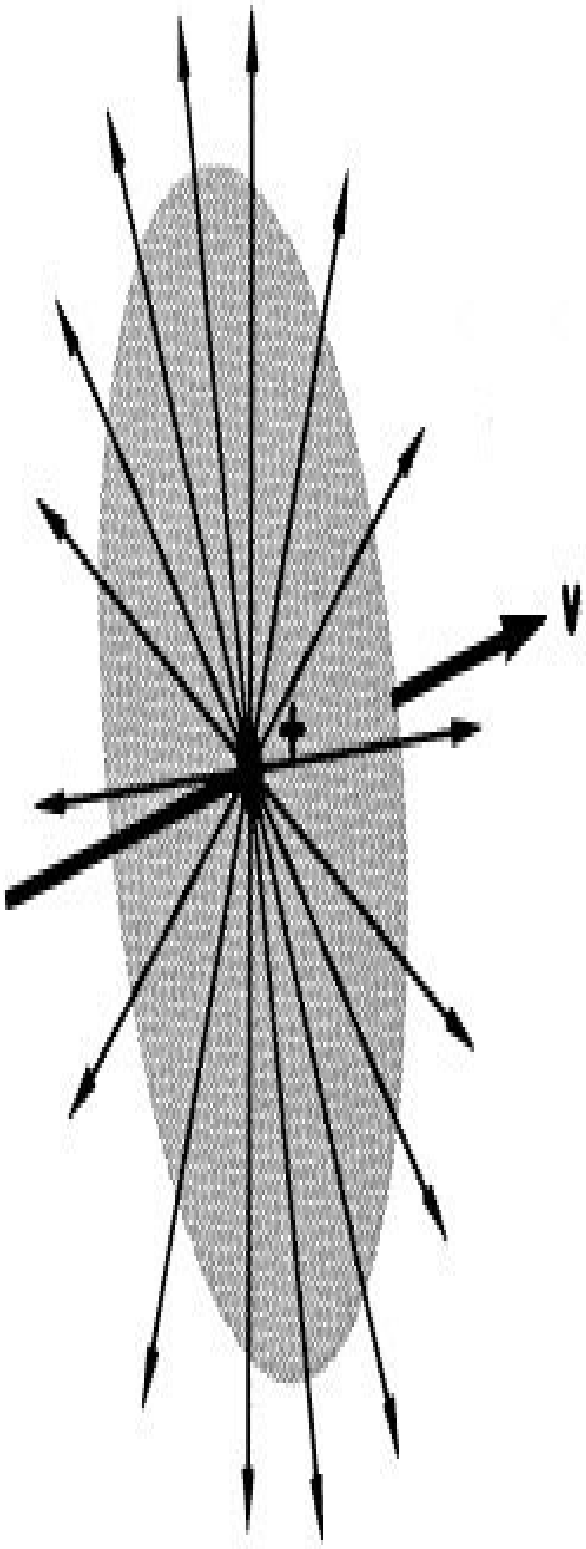
(14.16)

The electric field lines are still straight lines radiating from the charge, but the field pattern has been squashed in the direction of motion. At speeds close to c, when γ is large, the field on the x-axis is directed along the x-axis and has the magnitude

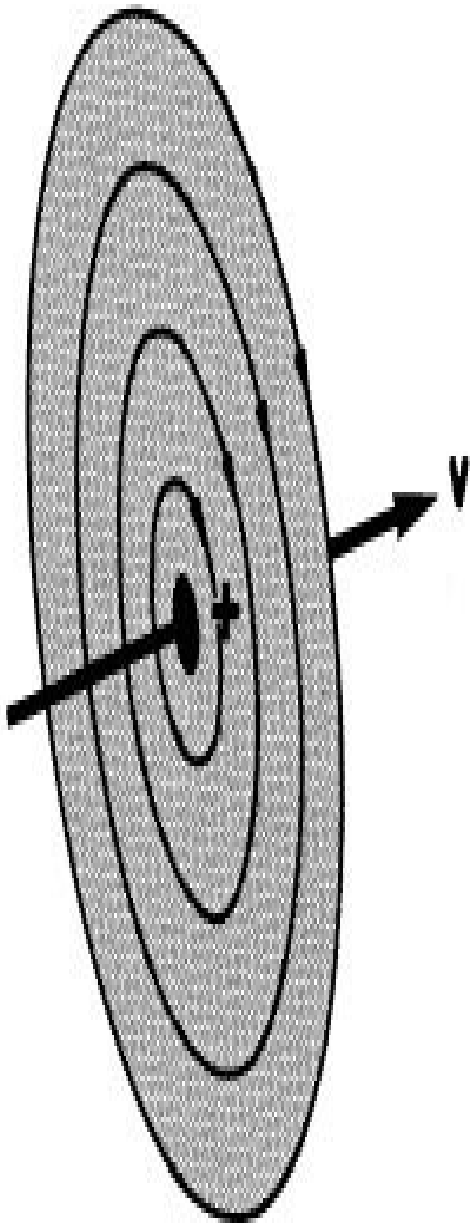


the same inverse square behaviour as for a stationary charge at $(x - vt)$ but diminished by the factor $1/\gamma^2$. On the other hand, the field radial to the charge at $x = vt$ is increased by the factor γ. Thus for very fast charged particles, such as the π-mesons in the experiment mentioned at the beginning of this chapter, for which $v/c > 0.99975$, the electric field is confined to a thin disc perpendicular to the direction of motion, as illustrated in Figure 14.4. Lines of B circulate around the moving charge, and Equation (14.16) shows that they too are confined to the disc where the electric field is appreciable.

Figure 14.4. Around a charged particle moving at a speed near to c, the fields are confined to a thin disc which moves along with the particle.



Lines of **E**



Lines of **B**

Transformation of forces

Look again at the transformation of the electric field from S to S′ in Equations (14.13). The right-hand side is reminiscent of the components of the Lorentz force on a charged particle moving at a speed v in the x-direction. This is no coincidence. In the S′ frame such a particle is at rest, and the whole of the Lorentz force must be accounted for by the electrostatic force. From Equation (14.13) the force on a charge q at rest in the S′ frame has components

$$F'_x = qE'_x = qE_x$$

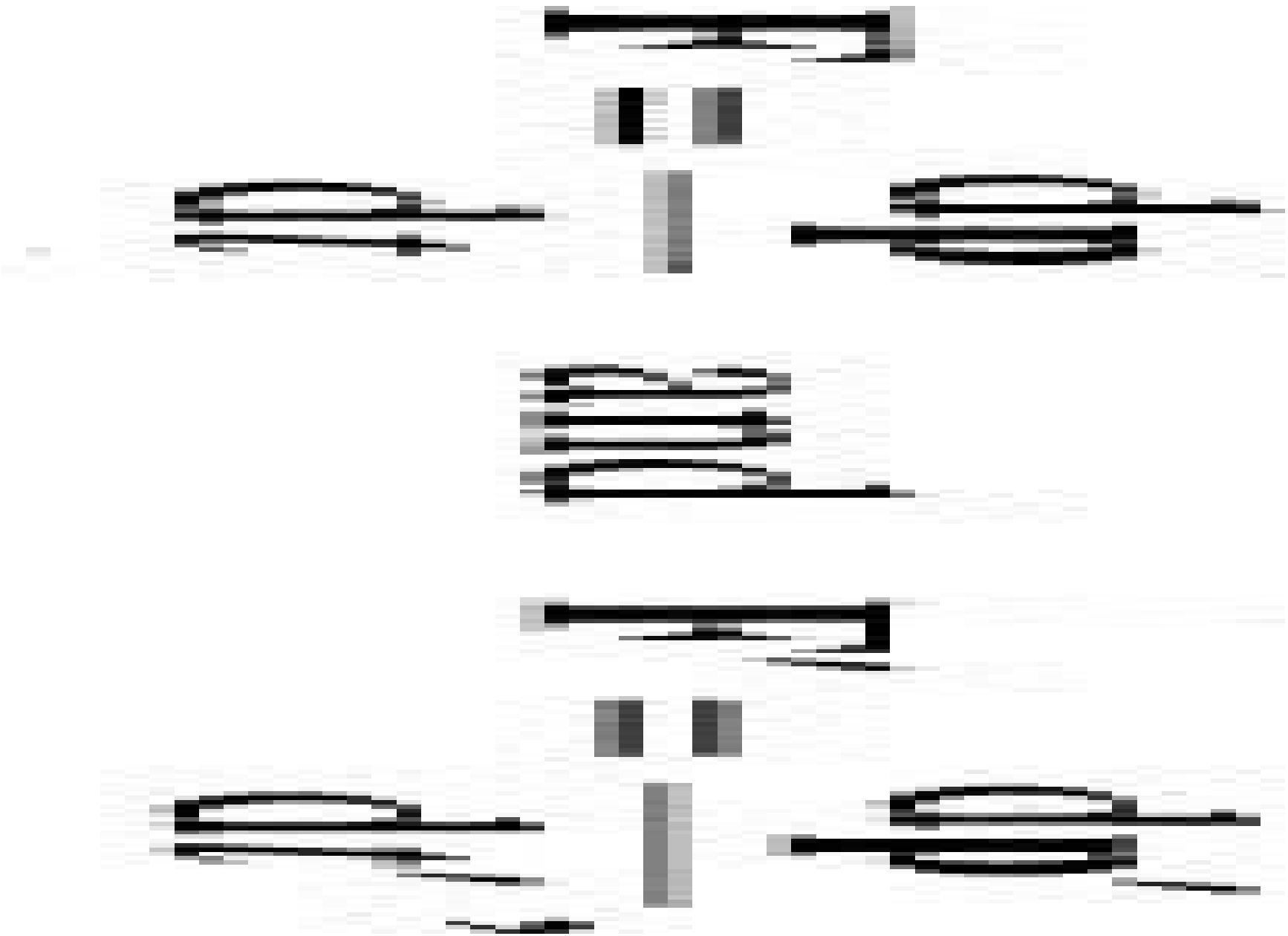
$$F'_y = qE'_y = q\gamma\left(E_y - vB_z\right),$$

$$F'_z = qE'_z = q\gamma\left(E_z + vB_y\right).$$

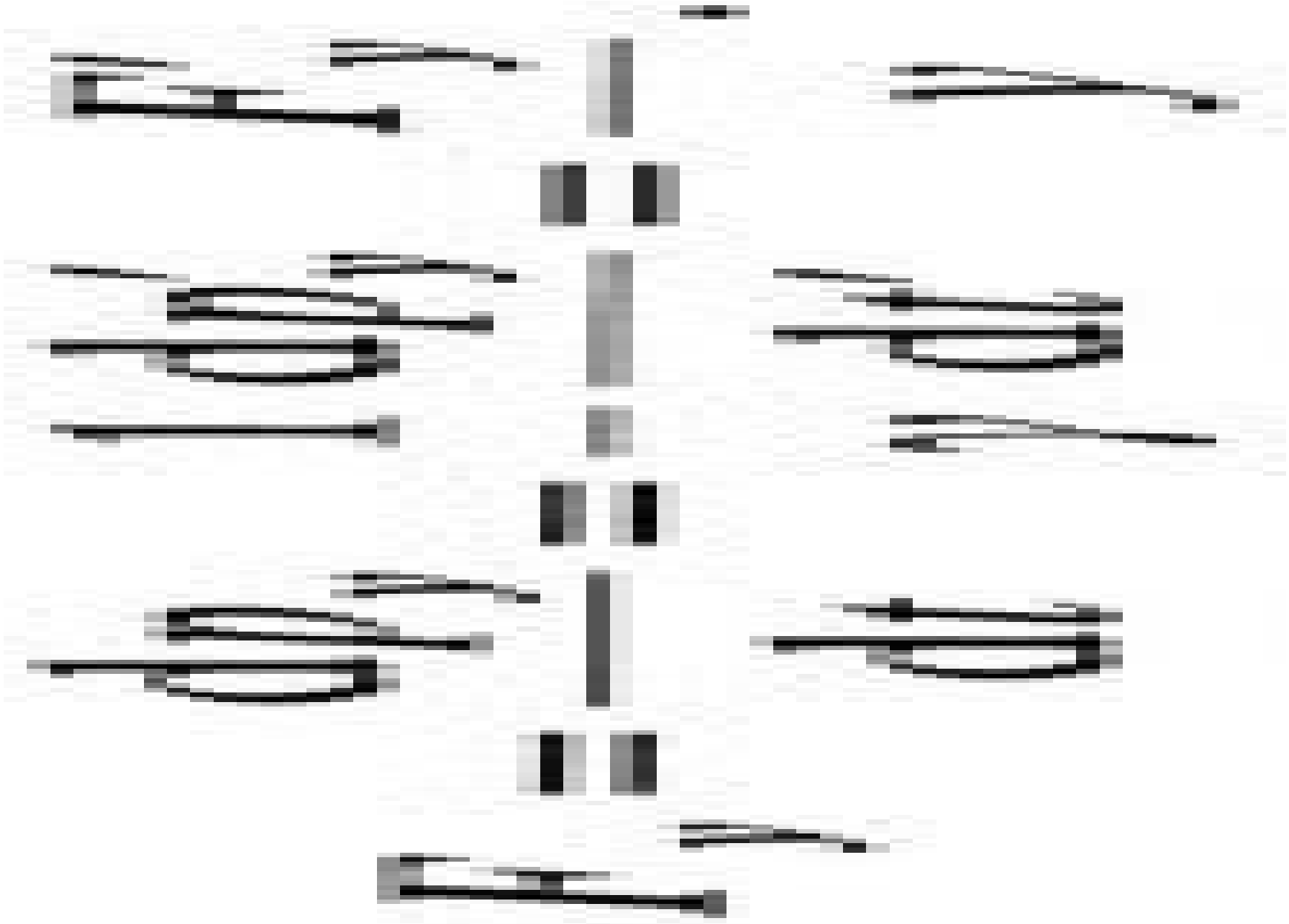
By comparison, in the S frame the components of the Lorentz force $F = q(E + v \times B)$ are

(14.17)

Is this the right way to transform forces between the frames S and S′? In order to be consistent, force must equal rate of change of momentum in both frames, i.e.



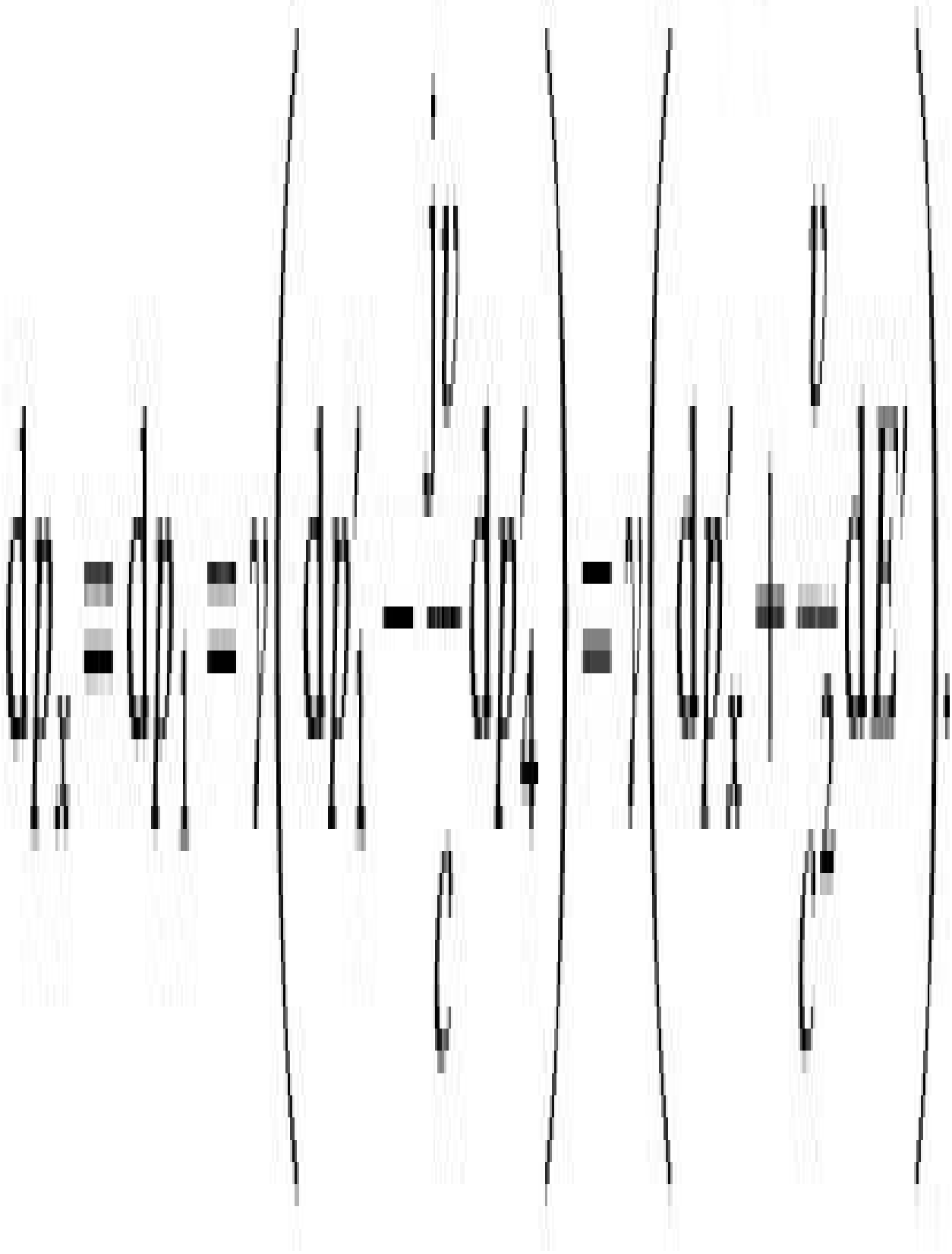
where the time derivative applies to a point fixed relative to the particle. An observer in S thinks that a clock moving with the particle runs slow by the factor γ, and thus $dt = \gamma\, dt'$. The increment dp_y of the transverse momentum component is unaltered by the Lorentz transformation, so that $dp_y = dp'_y$. The y-component of the force therefore transforms as follows:



Similarly

(14.18)

The transformation of the x-component of the energy-momentum four-vector (p, jE/c) is given by Equation (14.8) to be



Hence

(14.19)

since in the S′ frame the particle is momentarily at rest, and dE′/dt′ = 0. These results are consistent with the laws of transformation of force derived from the Lorentz force given in Equation (14.17).

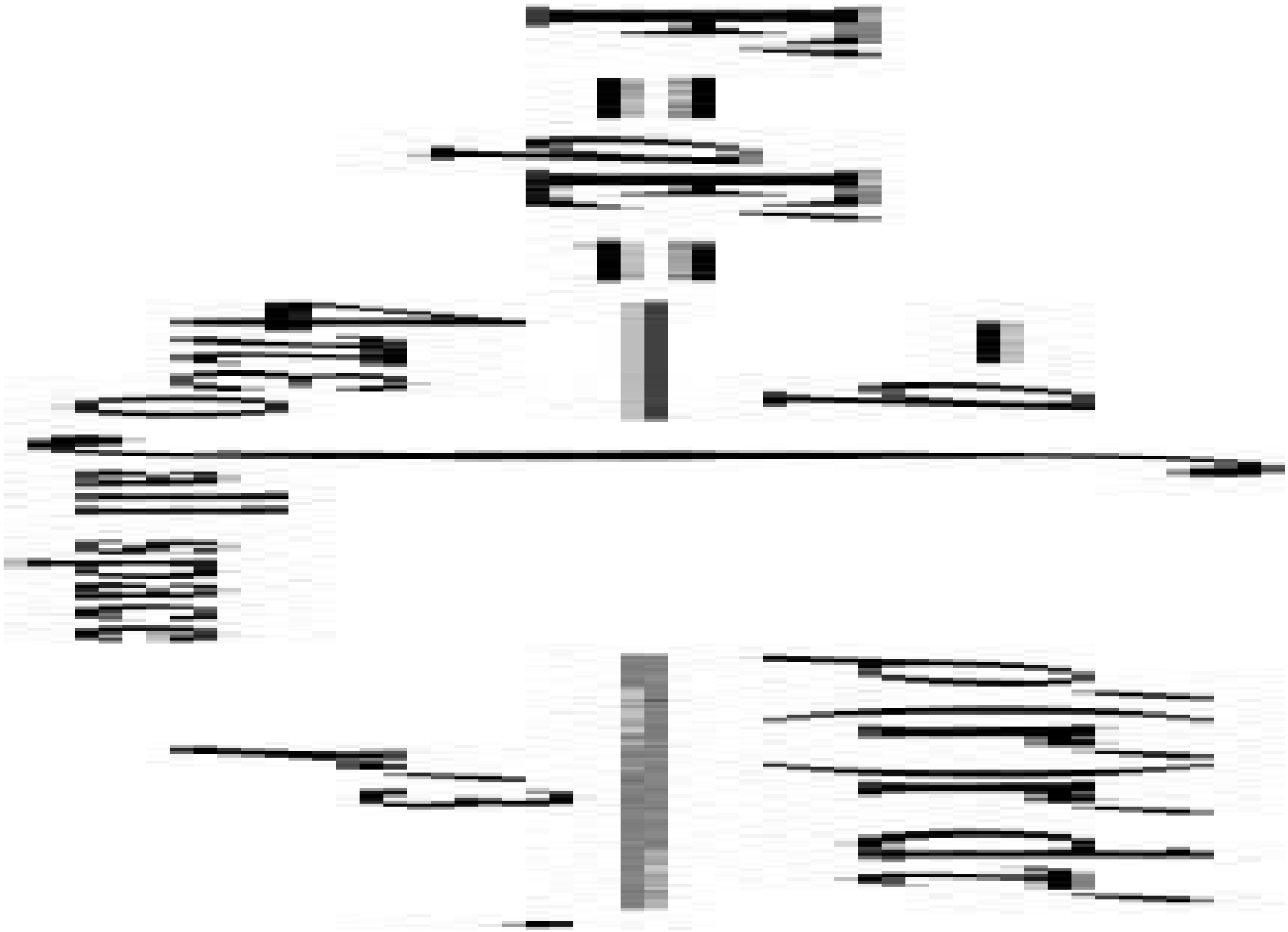
This completes the demonstration that the classical laws of electromagnetism are consistent with the special theory of relativity. We have been able to express Maxwell’s equations for the potentials as a single four-vector equation, so ensuring that this equation has the same form in all inertial frames. Finally, the Lorentz force exerted by electromagnetic fields on charged particles transforms in the appropriate way to agree with the relativistic relations between force and momentum. Different observers therefore have consistent descriptions of the motion of charged particles in electromagnetic fields.

14.7 MAGNETISM AS A RELATIVISTIC PHENOMENON

In this section the argument of the rest of the chapter is reversed. Up to now we have accepted the laws of electromagnetism, and tested them against the theory of relativity. Here we take the view that any acceptable theory must fall in with the requirements of relativity, since relativity is well-established in other fields of physics besides electromagnetism. Electric charge is a quantity which fits very easily into relativity theory: there is rather direct evidence that charge is invariant and that it is conserved for all observers in inertial frames. Accepting charge invariance and conservation, what else do we need to know? If we start with Coulomb's law for the force between stationary charges, is it possible, by using arguments based on relativity, to deduce the laws of magnetism? The answer to this question is that although a rigorous deduction is not possible, it can be made very plausible that the laws embodied in Maxwell's equations represent the simplest conceivable generalization of Coulomb's law which is consistent with relativity.

Let us see how far we can go from the starting-point of a distribution of stationary charges. As before, the charges are taken to be at rest in the frame S′. Applying Coulomb's law and the principle of superposition, the force on a test charge q at the origin of S′ is given in terms of the charge density $\rho'(\mathbf{r}')$:

(14.20)

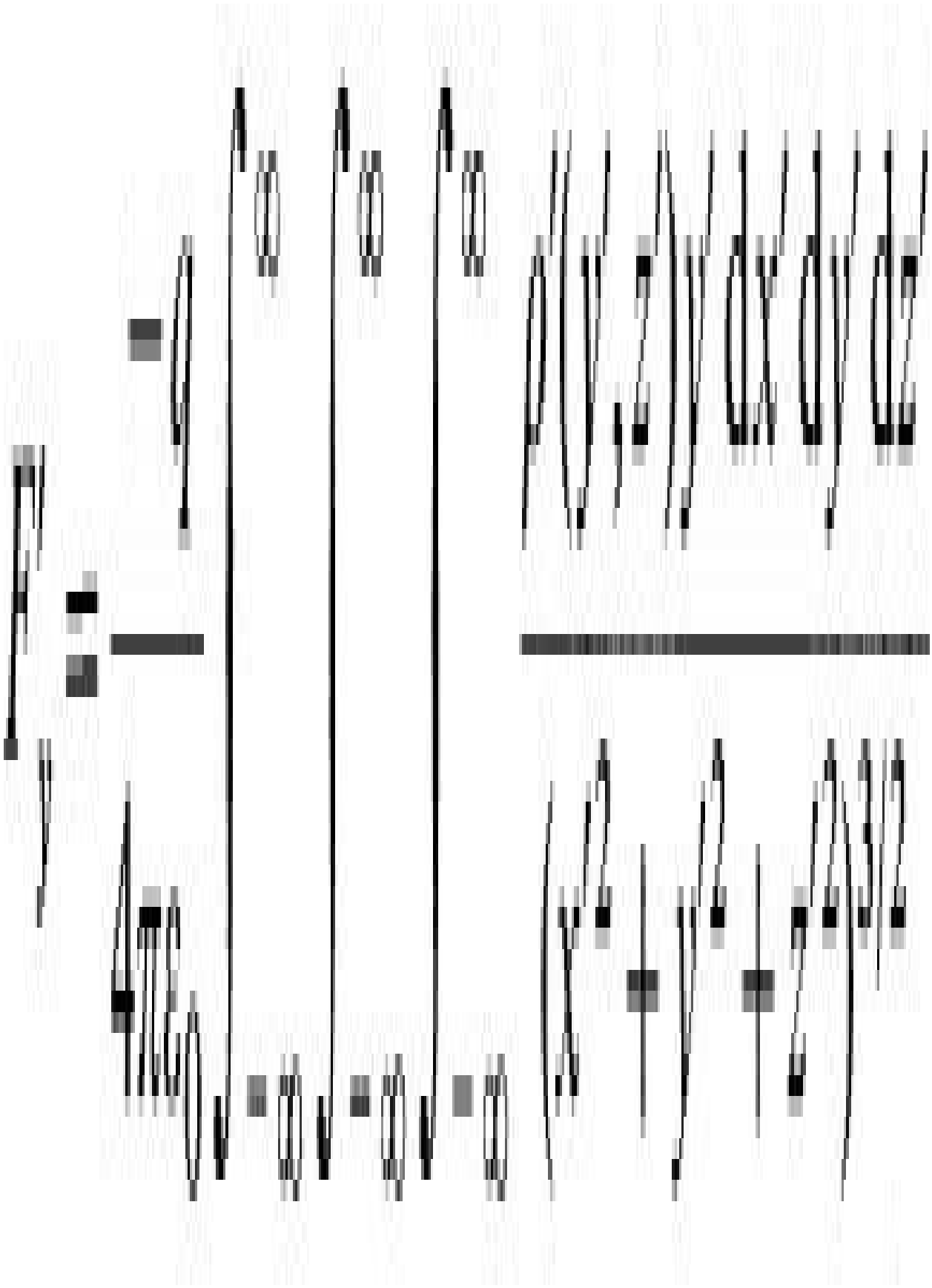


Here r′ is the vector from the origin to the volume element dτ′, as shown in Figure 14.5.

How is this situation described by an observer in a frame S, in which the charge is all moving at a speed v along the x-axis? From mechanics we know how to transform force (according to the transformations in Equation (14.18) and (14.19)) and since charge conservation requires that charge and current density form a four-vector (j,jcρ). we know how to transform ρ′. The transformation properties of each of the quantities in Equation (14.20) are thus known. Nevertheless it is not possible to transform the whole equation for the most general charge distribution. The trouble arises because the charge density at a fixed point in S may not be constant, and the integral on the right-hand side of the transformed equation cannot be evaluated without making additional assumptions. However, if the charge density $\rho'(r') = \rho'(y', z')$ as measured in S′ does not vary with x′, then ρ(r) is independent of time in S. In this special case, we can complete the transformation and look for a way of expressing the result in which j and jcρ are treated on an equal footing (the search is made considerably easier by the fact that we know the answer beforehand!). Then we may be able to discover the four-vector analogue of Coulomb's law.

Because of the uniformity along the x-direction, the x-component of the force on the test charge is zero in both S and S′ frames. In S′ the y- and z-components are given by

(14.21)



and

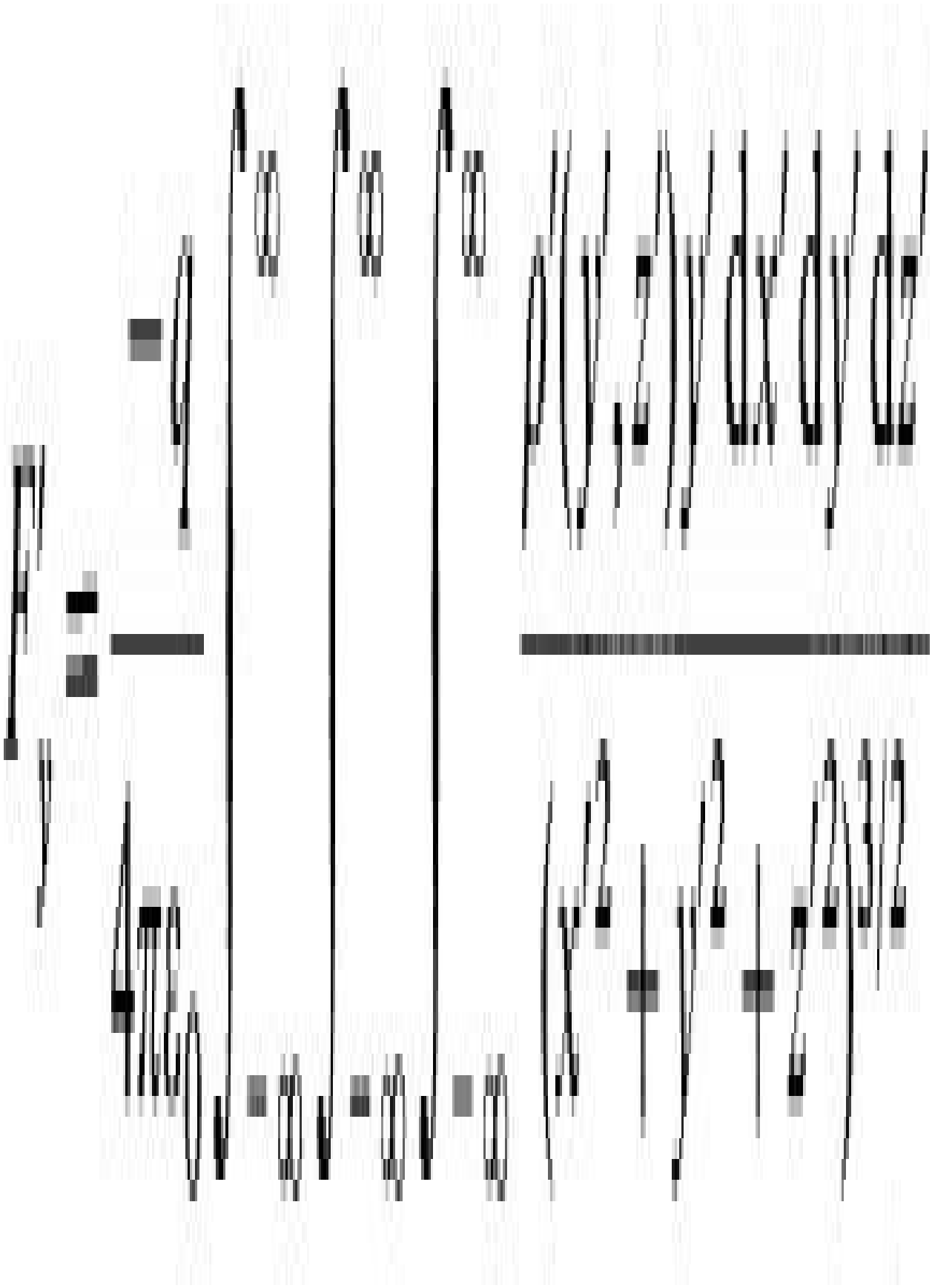
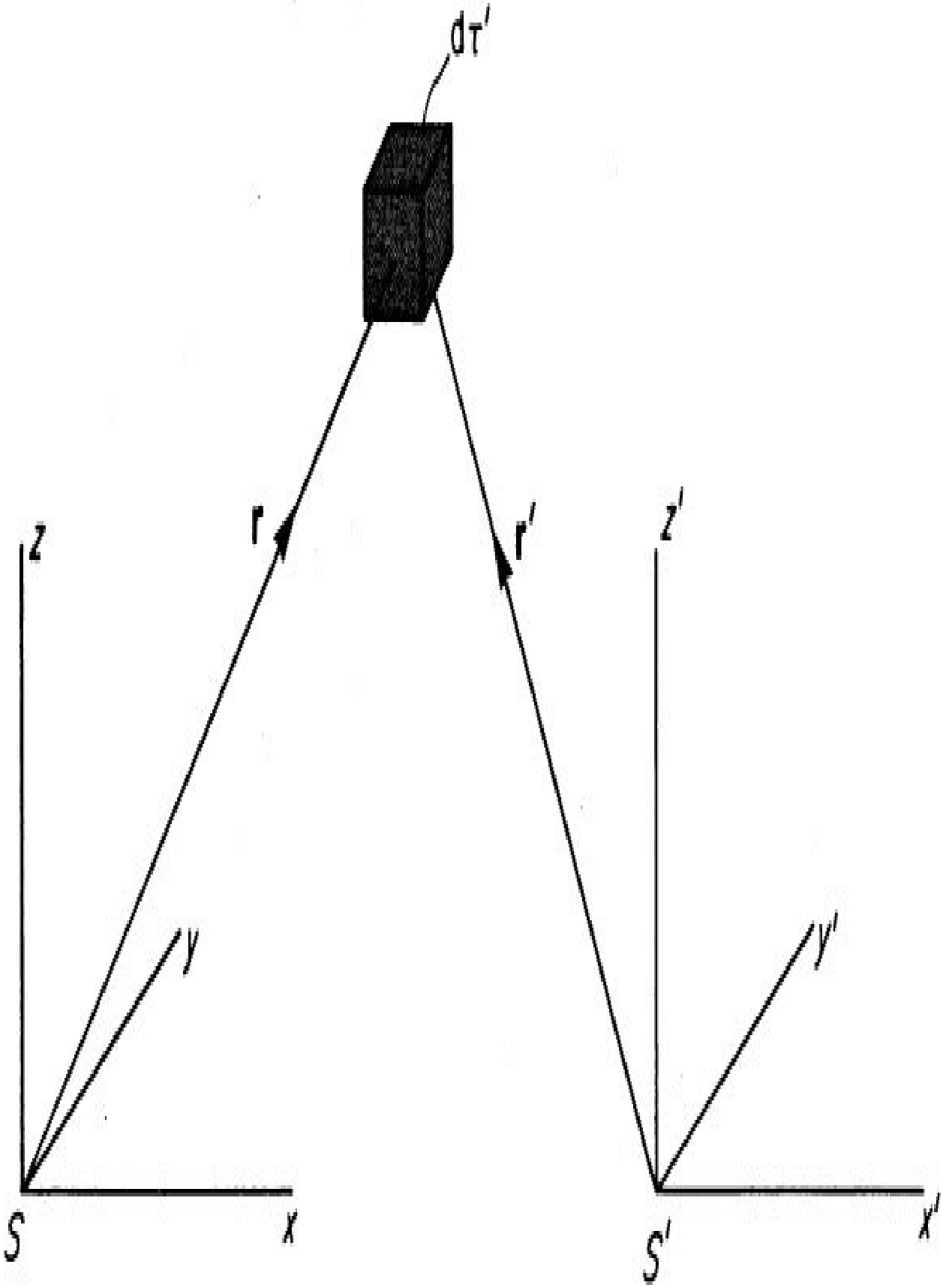
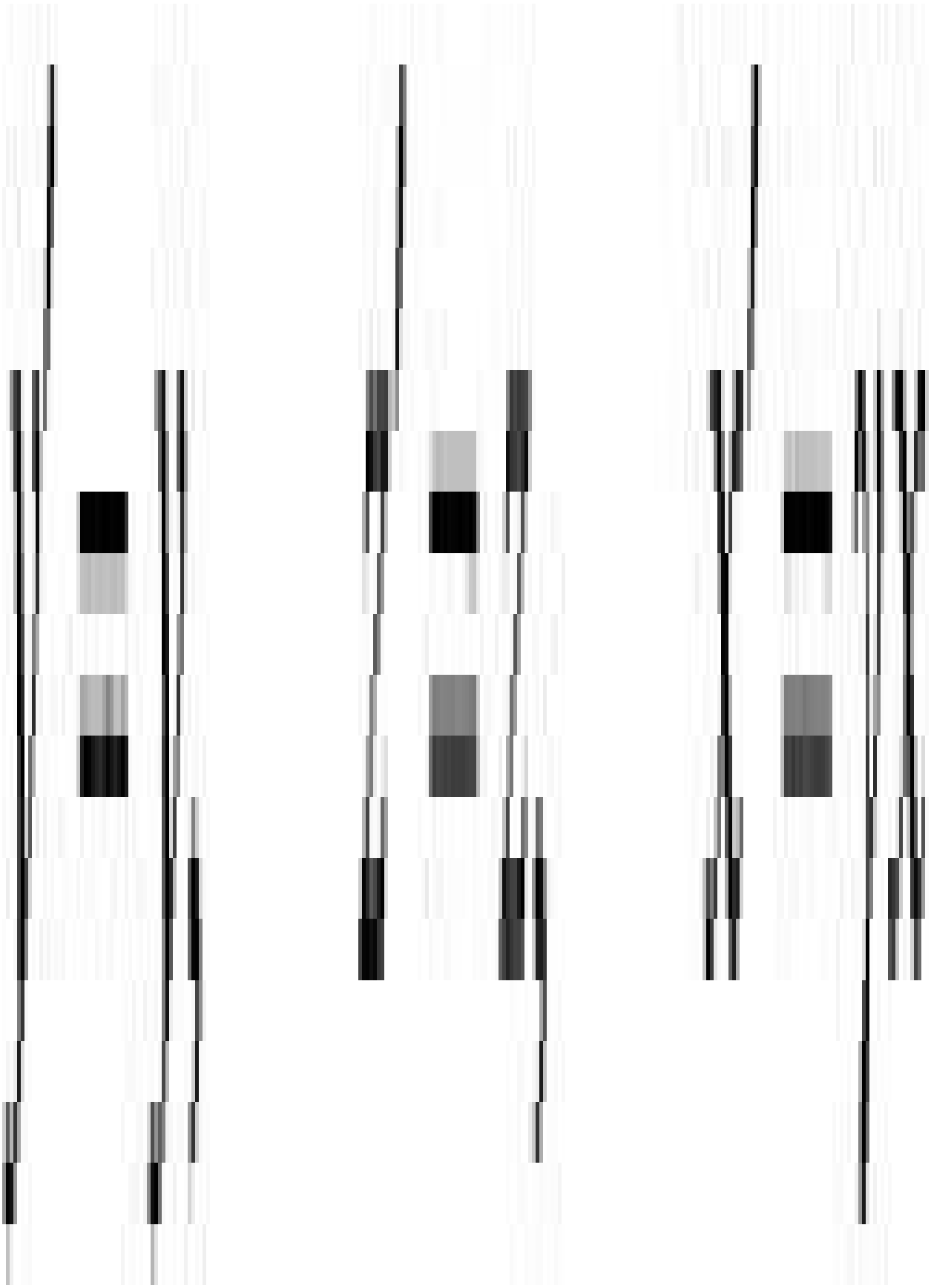


Figure 14.5. A charge element at rest in S′ is moving along the x-axis at a speed v in S.

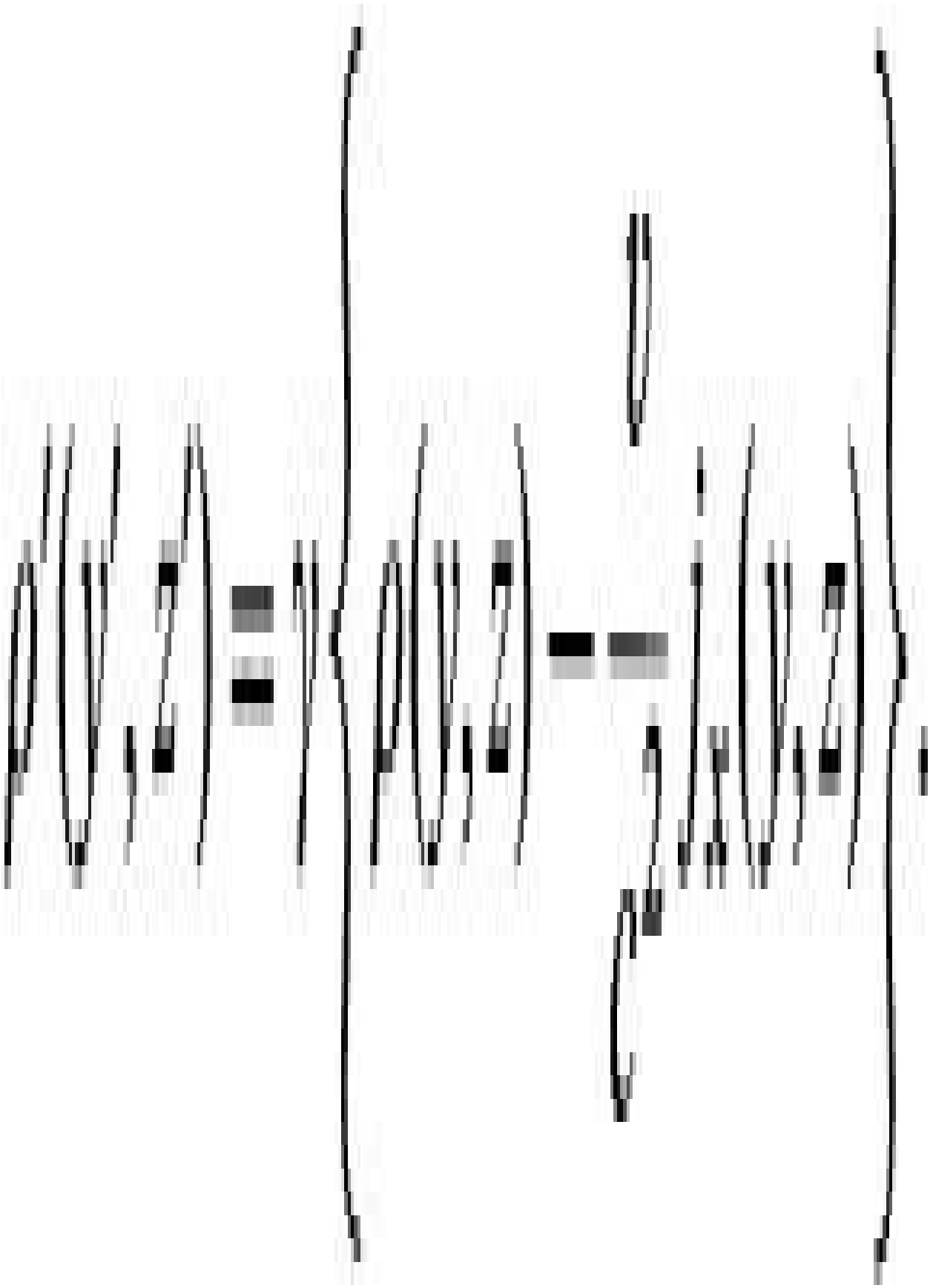

dτ'
r
r'
z
z'
y
y'
S
x
S'
x'

For simplicity we shall choose to measure the force on q at the moment when the origins of S and S′ coincide, i.e. at t = 0. The transformations needed for the y-component are then

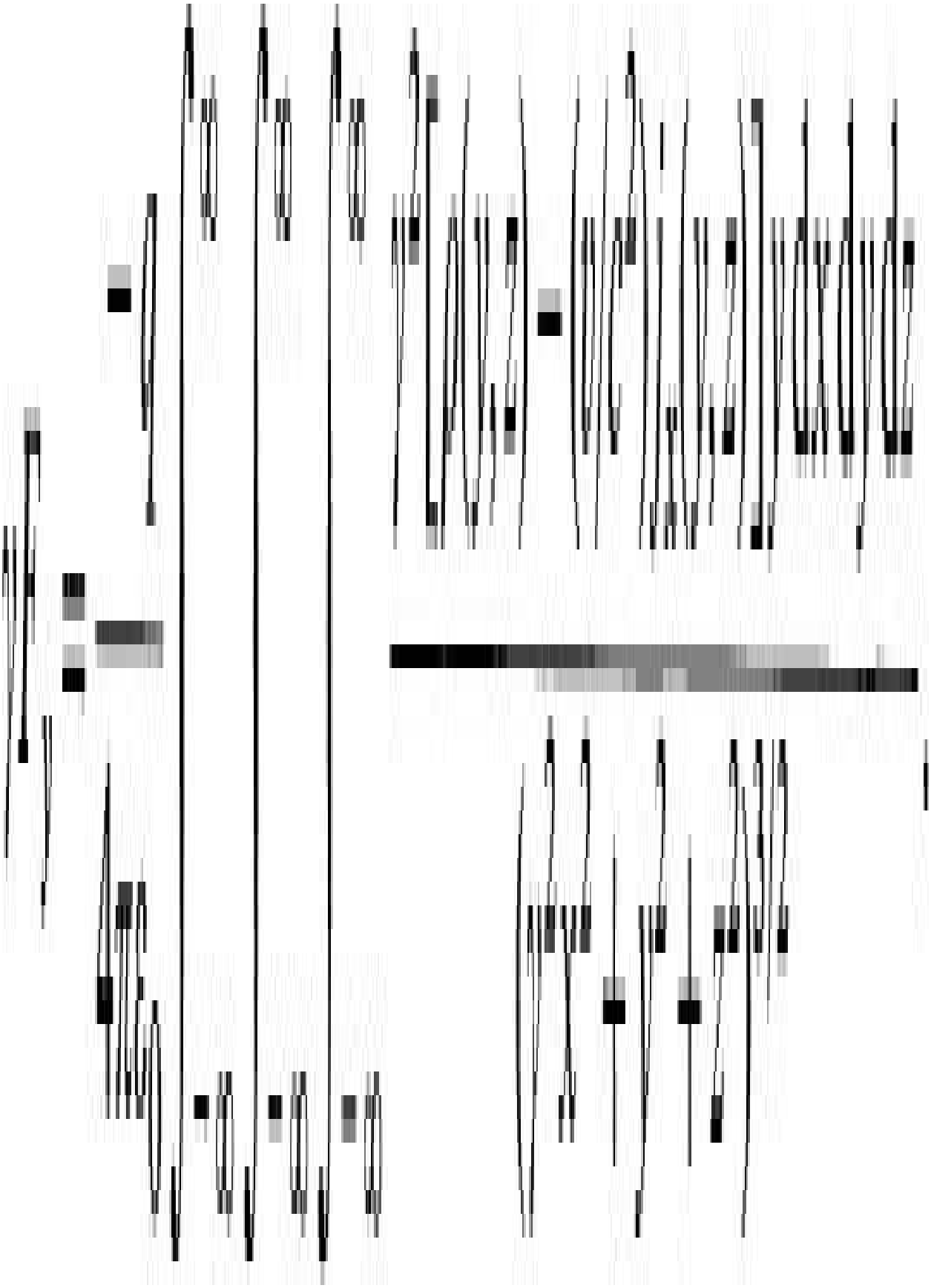
from Equation (14.18)



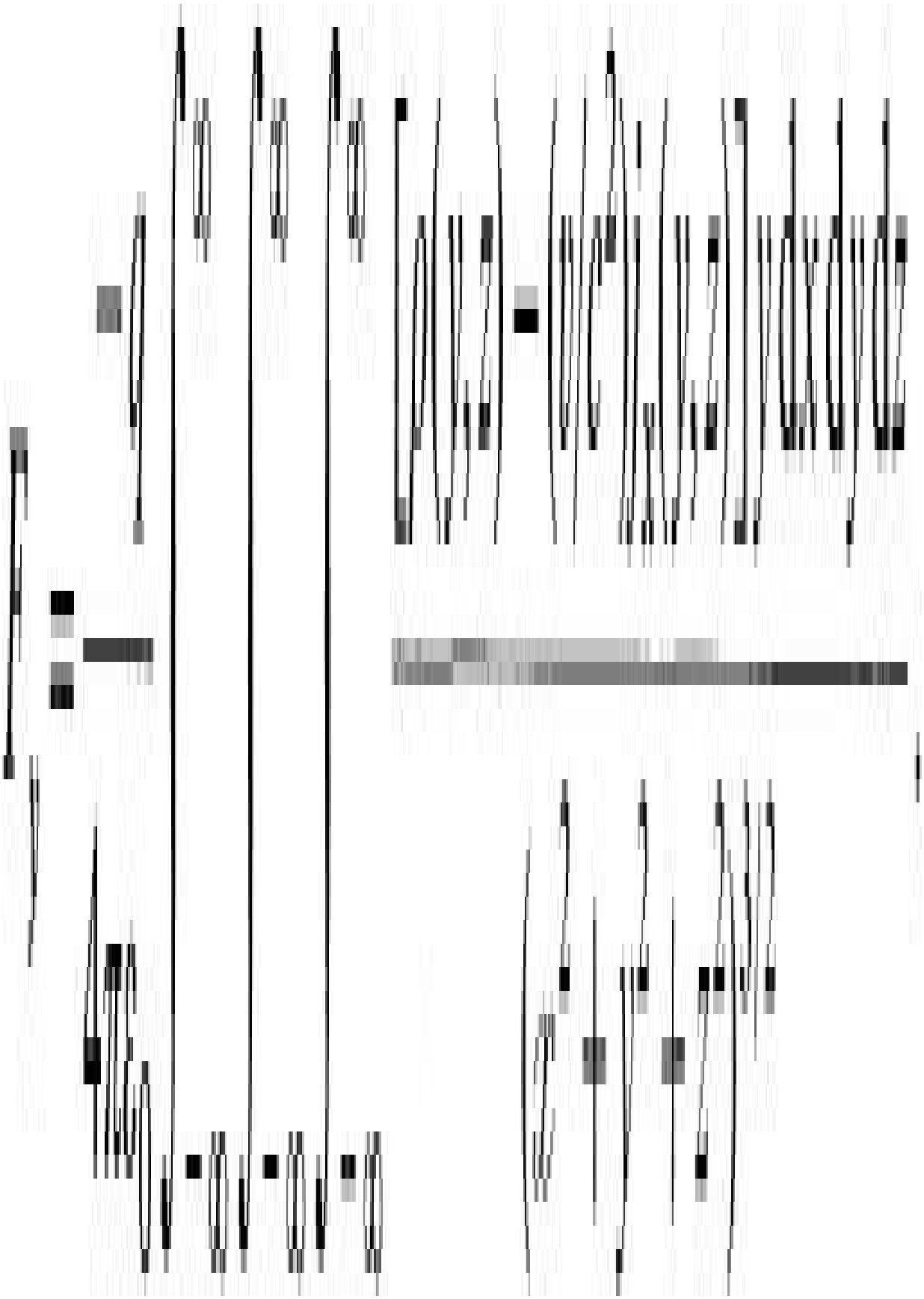
and



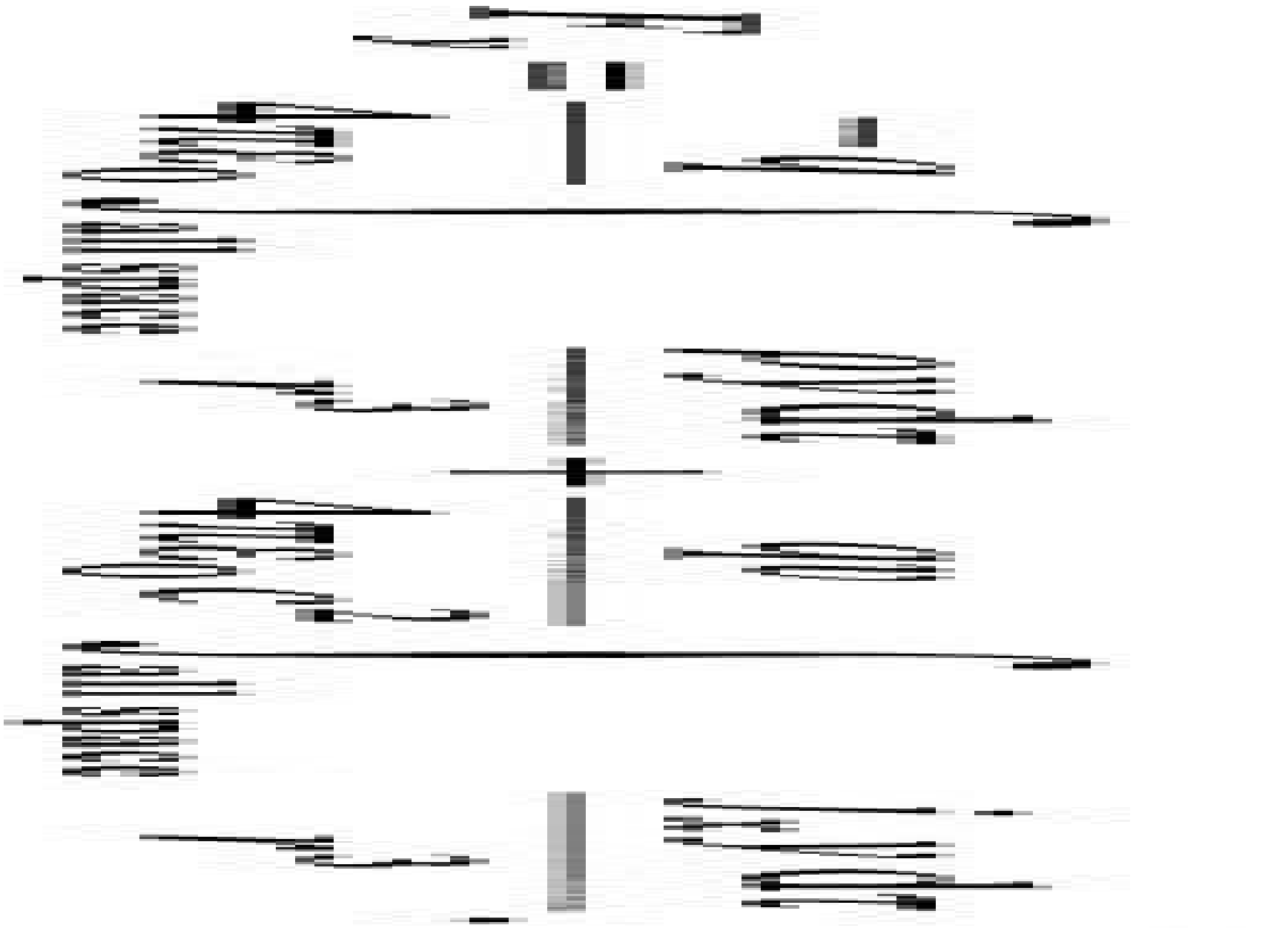
The charges are all moving in S, and an observer in this frame measures a current density which is everywhere in the x-direction; the y- and z-components of the current density are zero. Substituting in Equation (14.21) for the y-component of force, we find



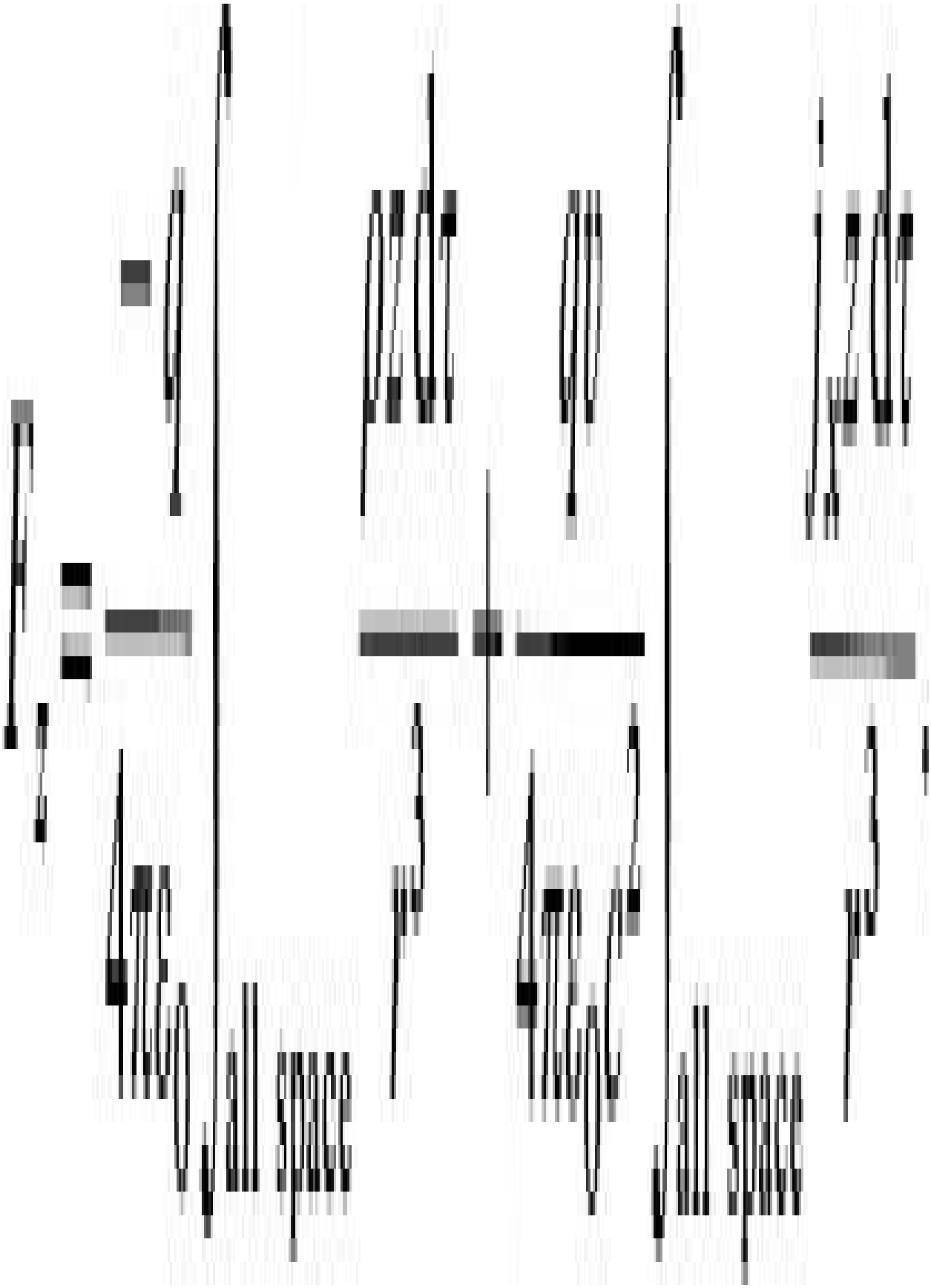
Introducing the new variable = γx



In this expression is simply a variable which is integrated from $-\infty$ to $+\infty$, and the result of the integral will be the same if we go back to our original set of variables and replace by x. The integral is now a volume integral in the S frame, which can be written

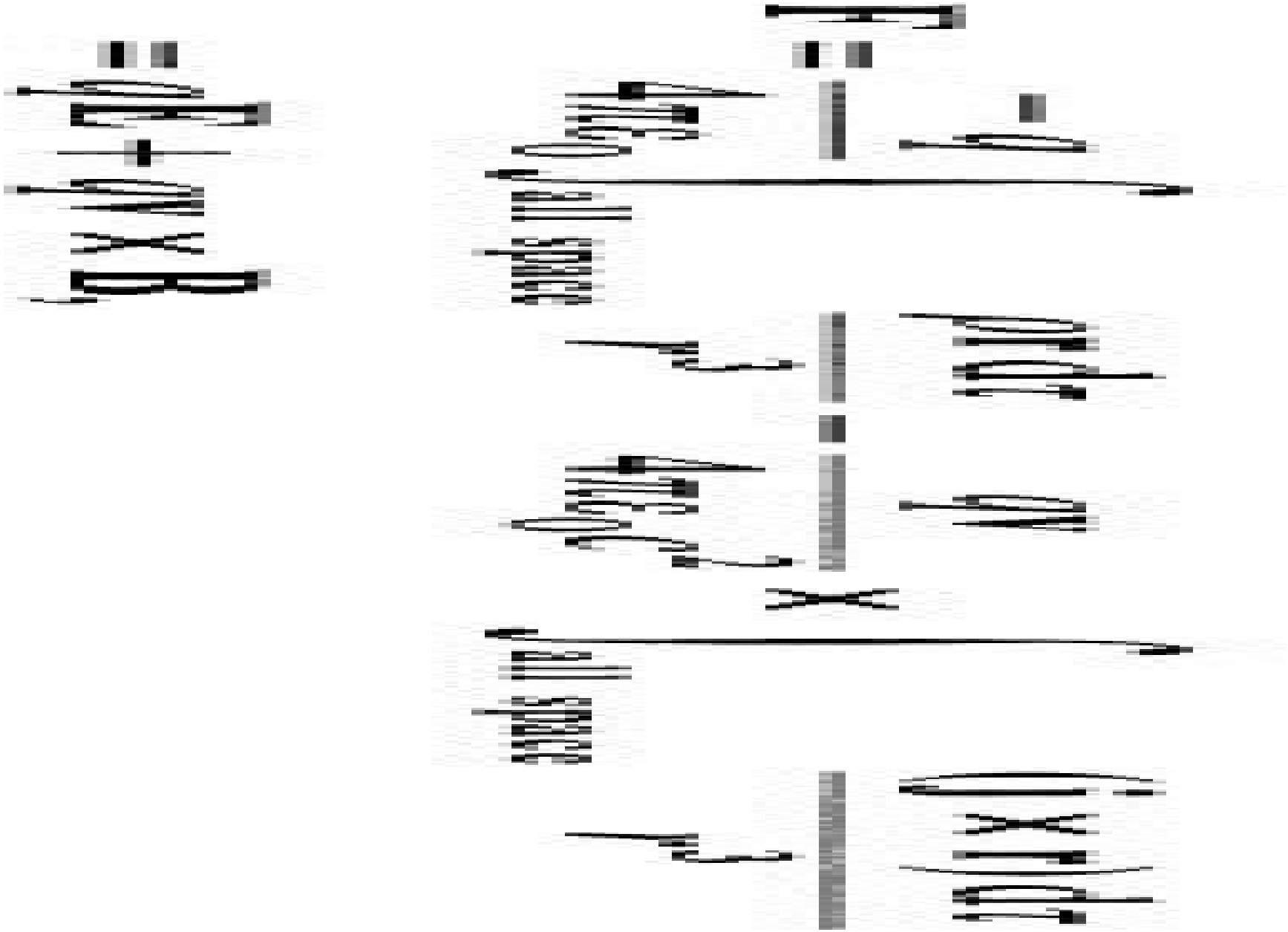


Similarly



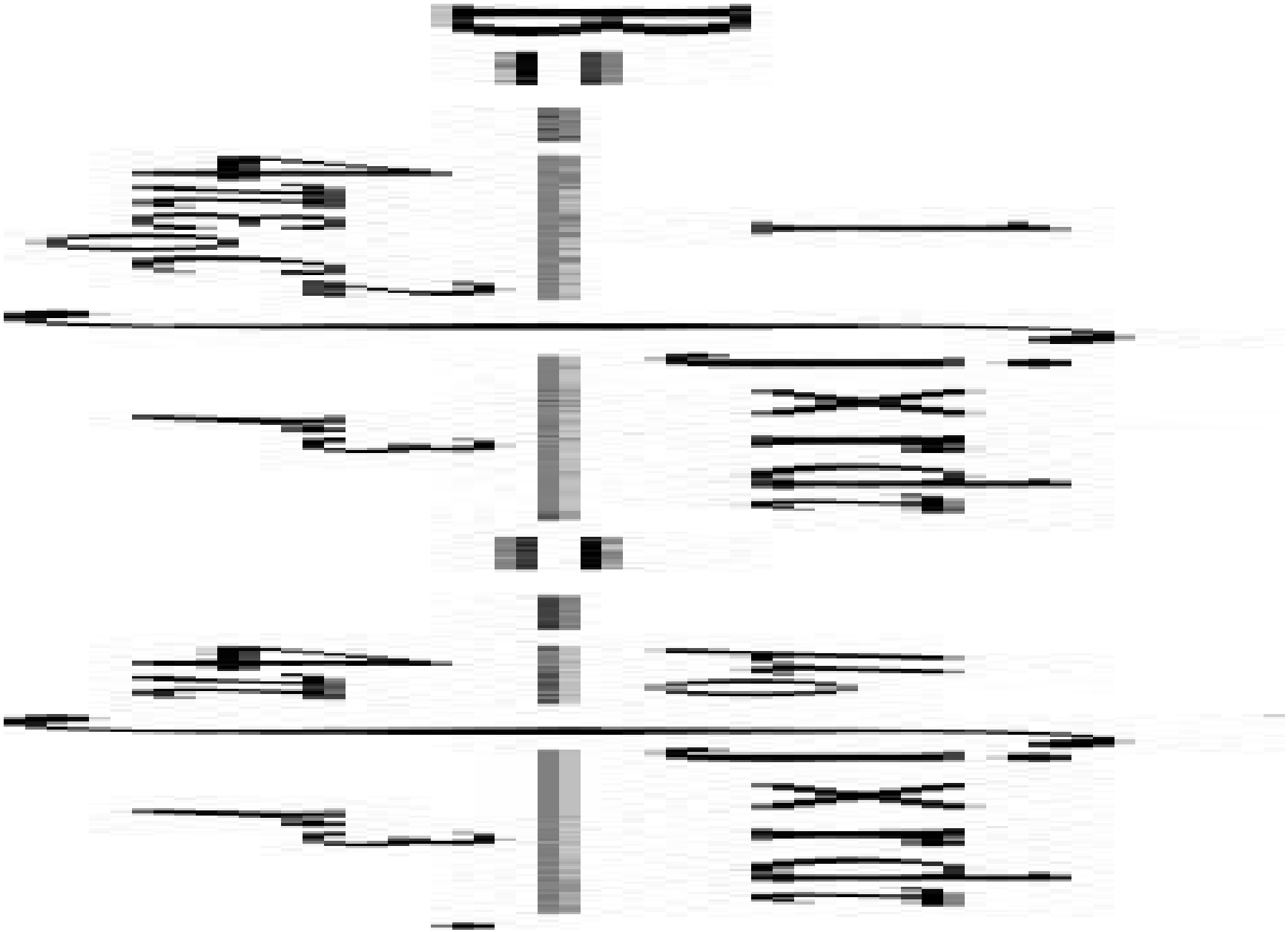
Remembering that Fx = 0 and that both j and v are pointing along the x-axis, the force can be expressed in vector notation as

(14.22)



writing

(14.23)

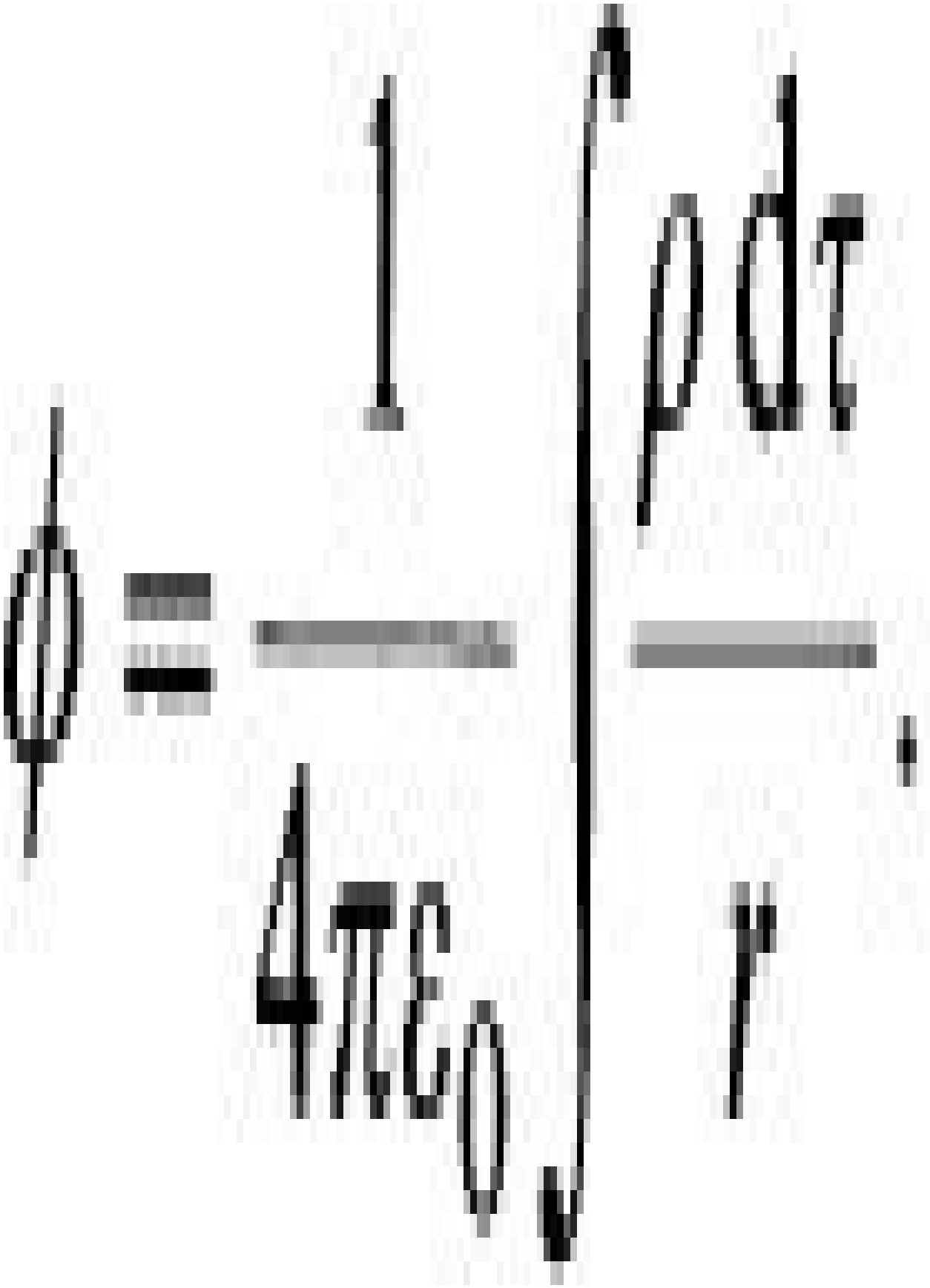


This expression is simply the Biot–Savart law* for the steady field B at the origin. We have thus derived the Biot–Savart law from Coulomb's law in the special case when there is a stationary charge distribution in the S′ frame, independent of x′. It is proved in Appendix C that the Biot–Savart law can be written in the equivalent form

(14.24)

where B = curl A.

Compare this equation with Equation (1.29) for the electrostatic potential ϕ; we have not included any surface charges in the present discussion and the potential reduces to



We see that A is related to j/c^2 in exactly the same way as ϕ is related to ρ, and we can construct a four-vector $(A, j\phi/c)$ which transforms in the same way as $(j, jc\rho)$. This four-vector is defined by Equations (14.24) and (1.29) for any distribution of charge and current density, and it is not restricted to charges which are all moving with the same velocity. However, after removing this restriction it is not possible to prove that the force on a charge q is still given by Equation (14.22). From the starting-point of Coulomb's law the Lorentz transformation cannot by itself tell us about the interaction of charges in relative motion. But let us now assume that Equation (14.22) holds in general, with B = curl A. Does the force then have the correct properties when it undergoes a Lorentz transformation? The answer to this question, which was discussed in section 14.6, is yes, provided that a time-dependent term is added to the electric field, which must become $E = -\text{grad}\,\phi - \partial A/\partial t$. This is Faraday's law, since

$$\operatorname{curl} \mathbf{E} = -\operatorname{curl} \operatorname{grad} \phi - \frac{\partial}{\partial t}(\operatorname{curl} \mathbf{A})$$

$$= -\frac{\partial \mathbf{B}}{\partial t}$$

To sum up, after making a Lorentz transformation of electrostatic forces, a simple generalization leads very persuasively to the other laws governing the forces between charged particles. There is a sense, then, in which Coulomb's law can be regarded as fundamental, with magnetic forces and induced e.m.f.s appearing as relativistic side-effects. Yet one ought not to attach too much importance to the philosophical game of choosing the neatest set of initial postulates as the basis of electromagnetism. The real test of a physical theory comes in comparing its predictions with observation. On this criterion electromagnetism is extraordinarily successful; as we pointed out at the beginning of this book, when quantum effects have been taken into account, no accurate observation has ever detected the slightest deviation from the theoretical predictions.

14.8 RETARDED POTENTIALS FROM THE RELATIVISTIC STANDPOINT

A general solution of Maxwell's equations in the absence of dielectric and magnetic media was given in Chapter 13 in terms of retarded scalar and vector potentials. The potentials are called 'retarded' because the contribution of each element of charge and current is delayed by the time taken for light to travel to the point where the potential is being calculated: information about the present value of charge and current propagates outwards at a speed c. In this section the retarded potentials are derived in an elegant way using relativistic arguments. We start by using the Lorentz transformation to find the potentials generated by a small charge when it is viewed from a moving frame of reference. Then the potentials caused by any distribution of charge and current can be found by superposition.

A charge $\rho' \, d\tau' = \rho' \, dx' \, dy' \, dz'$ is stationary in a frame S′ and is located at the origin of S′ as in Figure 14.6. What are the potentials seen at a time t in the S frame, at the point P which is stationary in S? We shall choose the origin of S so that a light signal leaving the origin at t = 0 reaches P at the time t—this choice of origin simplifies the algebra. The arrival of the light signal at P is an event with coordinates (r, jct), where t = r/c. The origins of S and S′ coincide at t = 0, and for an observer in S′ the light signal arrives at P at the time $t' = r'/c$. In S′ the contributions to the scalar and vector potentials at P due to the volume element at the origin are

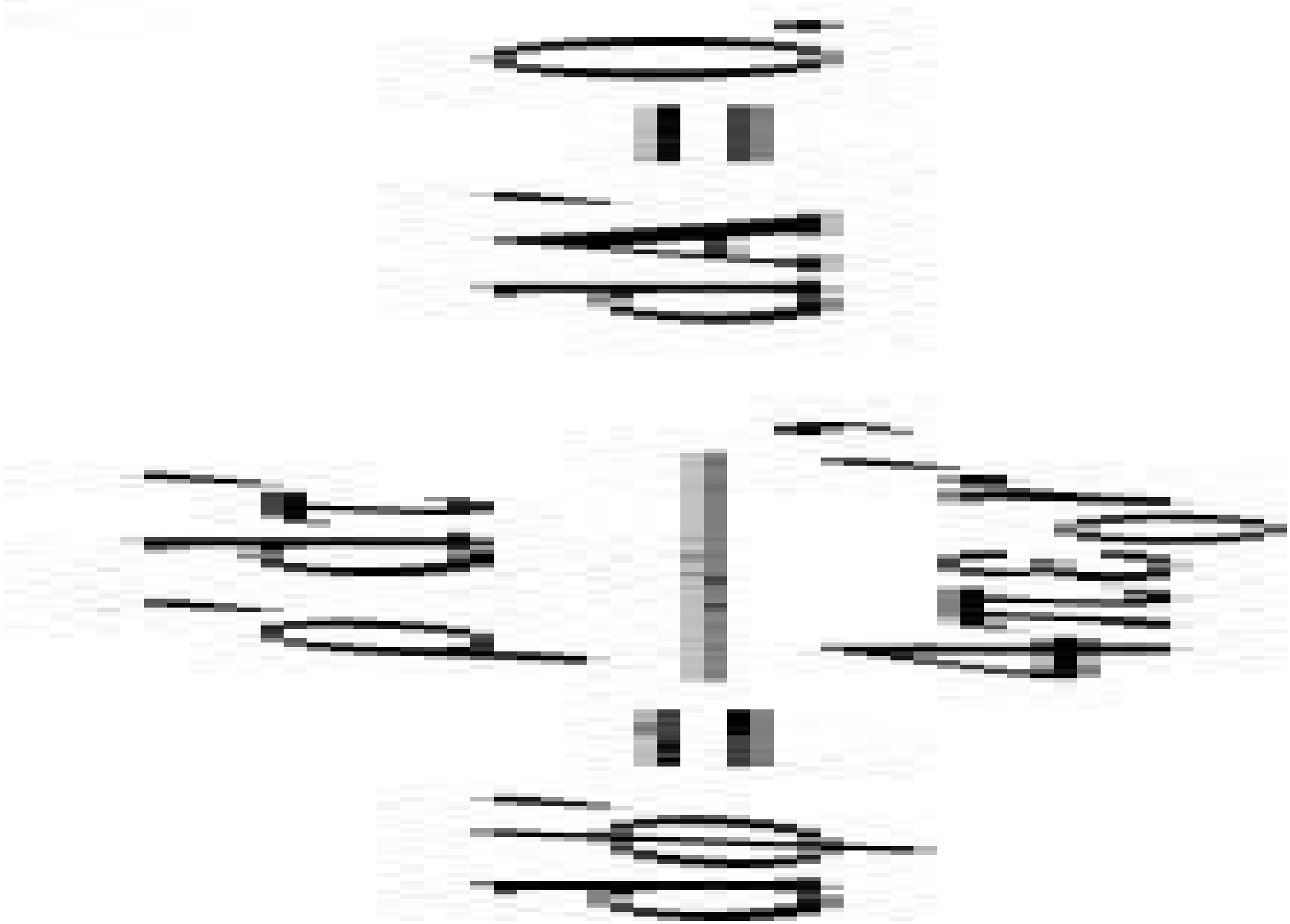
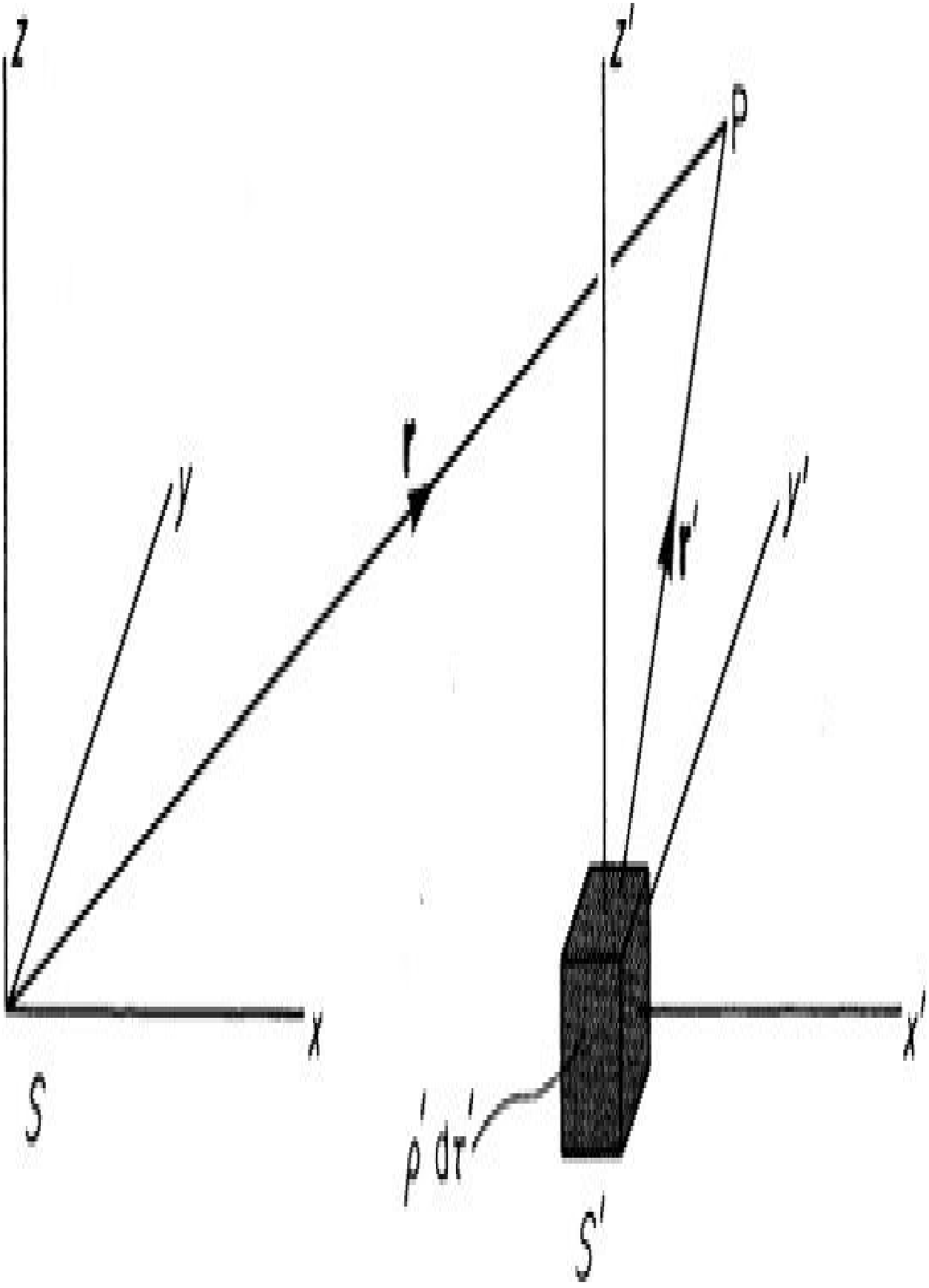


Figure 14.6. An element of charge at the origin of S′ moves past the point P which is at rest in S.


z
z'
P
r
r'
y
y'
x
x'
S
ρ' dτ'
S'

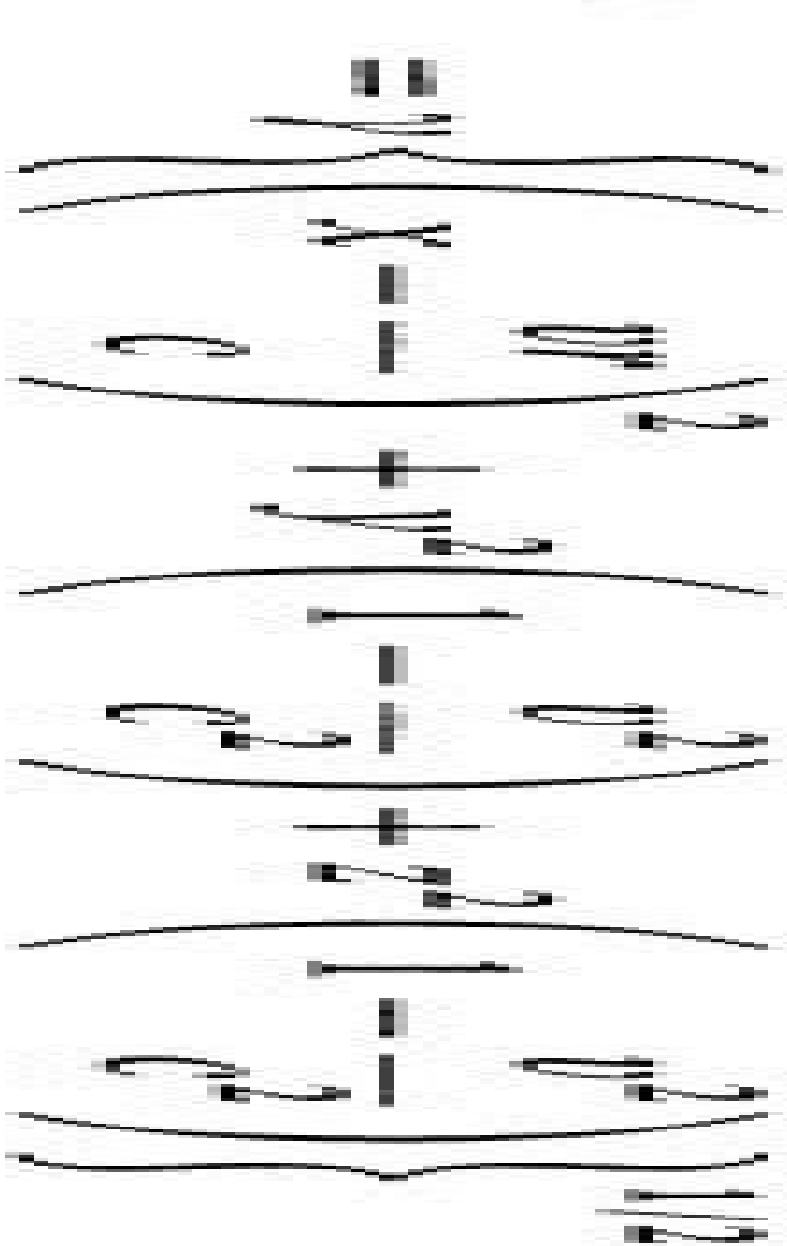
Knowing that (A, jϕ/c) is a four-vector, the potentials at P as seen in S can now be obtained from the Lorentz transformation:

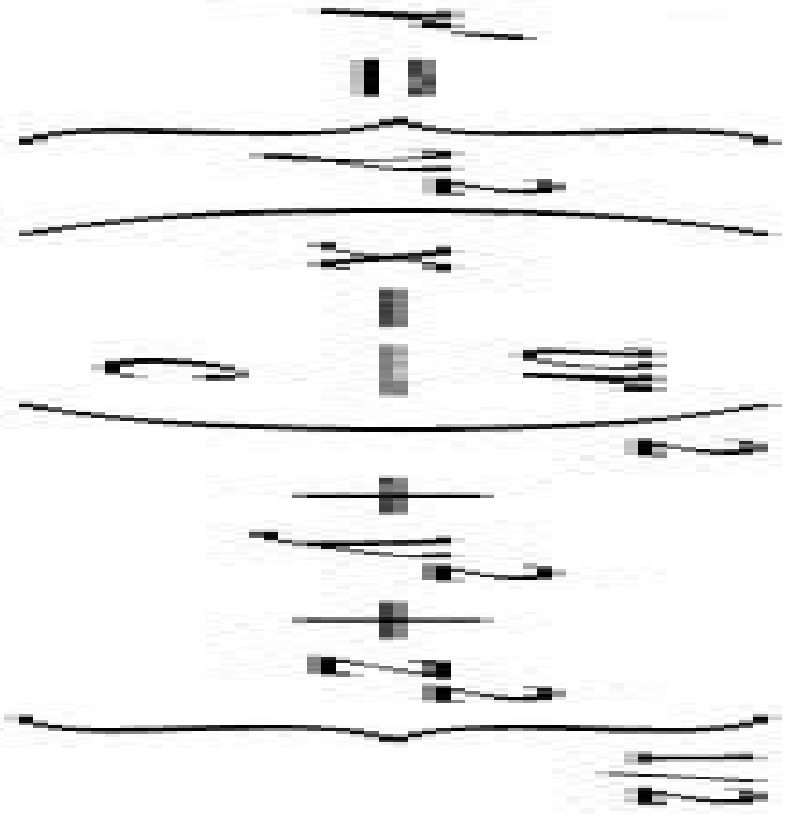
(14.24)

The right-hand sides of these equations have to be transformed back again to give the potentials in terms of the charge and current densities in S. The time in S is t = r/c, and we have

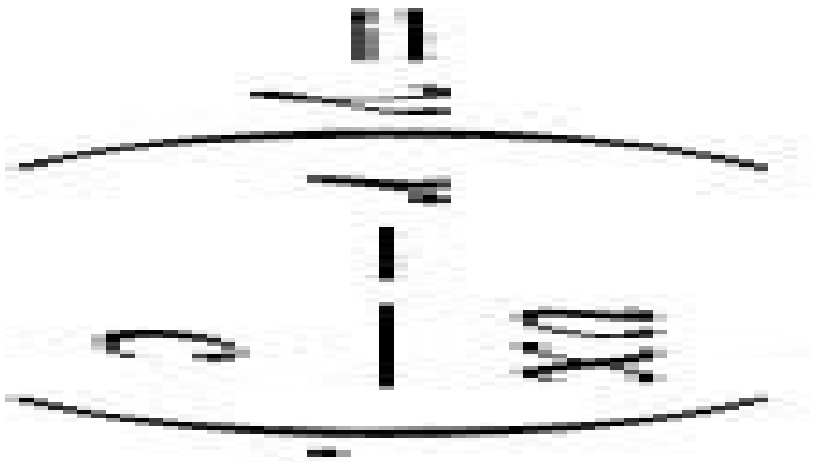
(14.25)

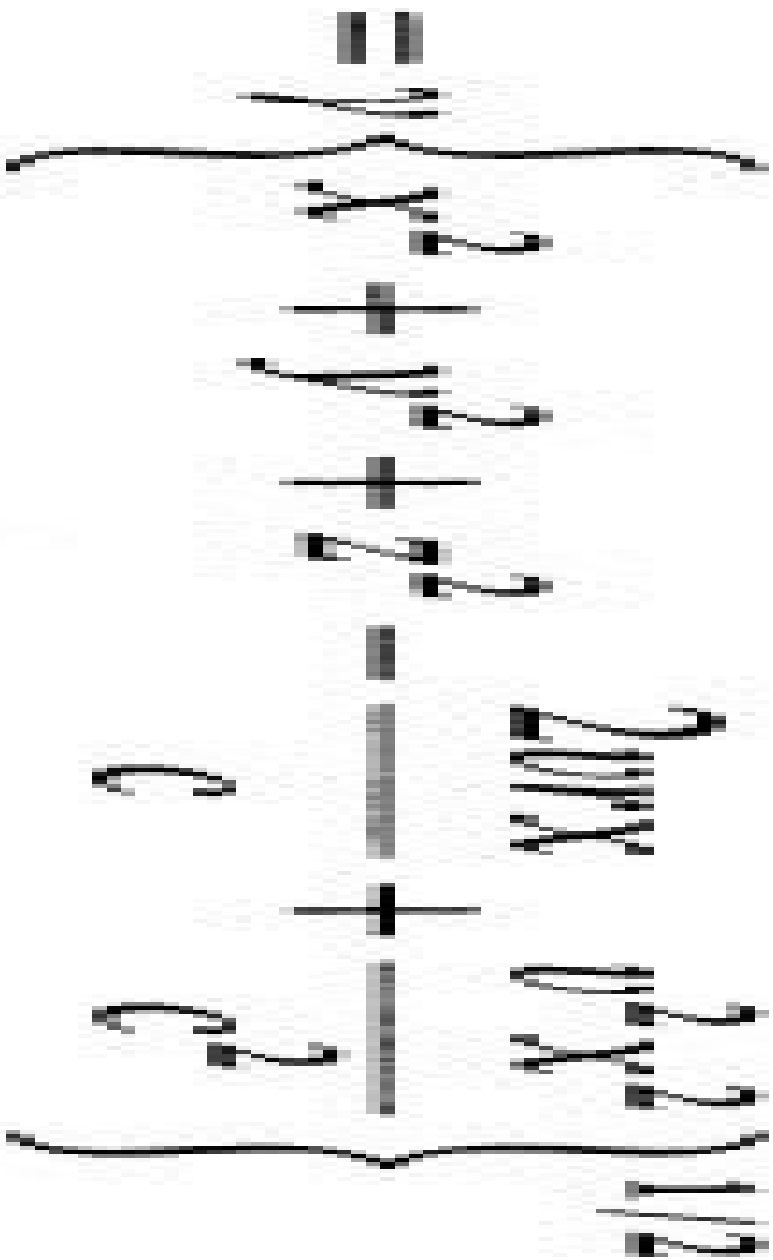
Hence





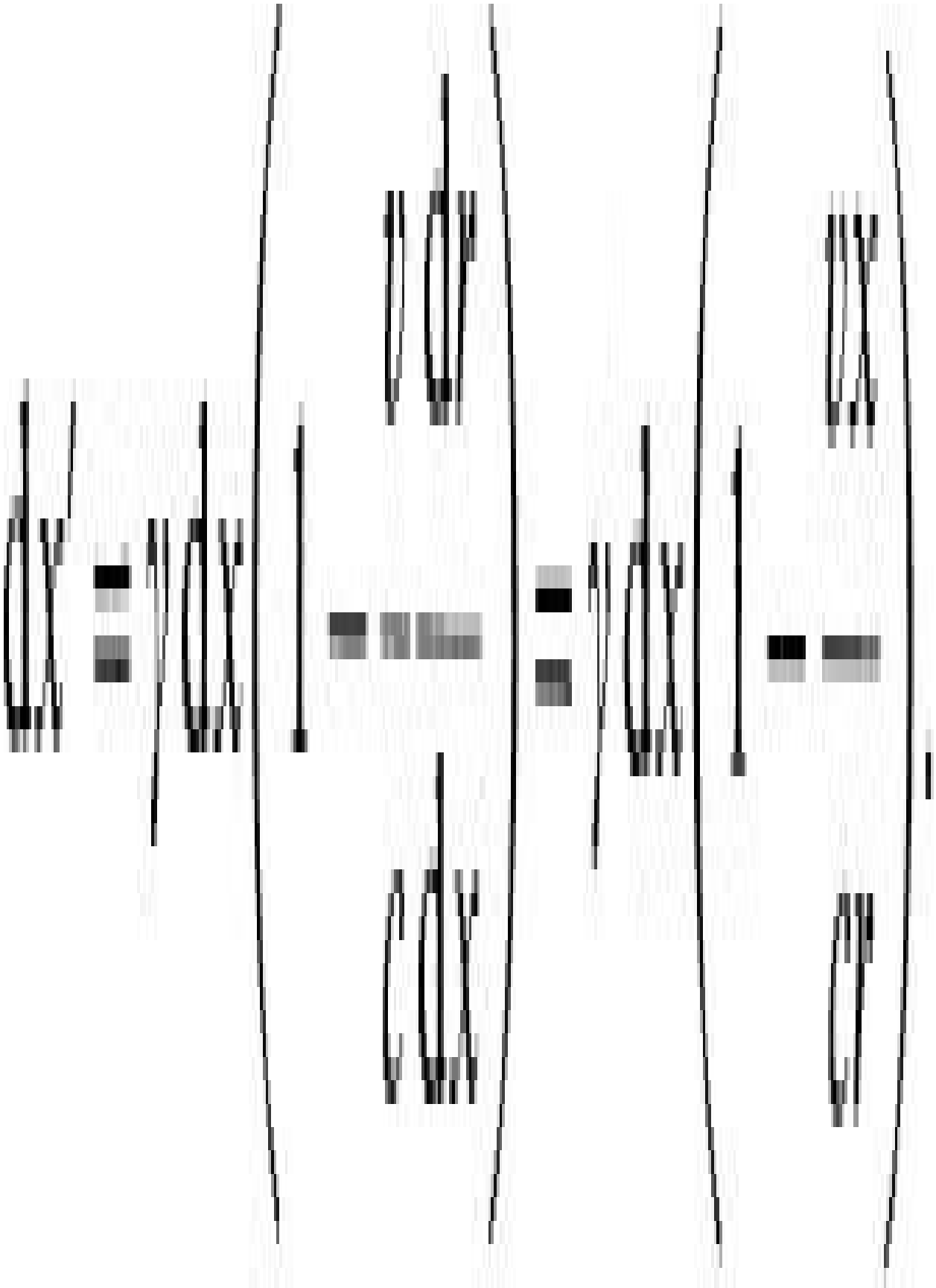
(14.26)





The transformation of $d\tau'$ is tricky, because the position of the origin of S varies as x' varies. Even across the distance dx' there is a difference in the time taken for a light signal to reach P, with the result that the length dx in S is stretched out by comparison with a frame having a fixed origin. This is a kind of shock-wave effect. As we have seen in section 14.6, the fields caused by a particle moving at a speed very close to c pile up into a thin disc travelling along with the particle, like the shock-front accompanying an aeroplane flying at a little less than the speed of sound. To account for this shock-wave effect properly we must remember that in Equation (14.25), r is a variable. Differentiating this equation, we find

(14.27)



As before $dy' = dy$ and $dz' = dz$. Finally, the transformation of charge density is

(14.28)

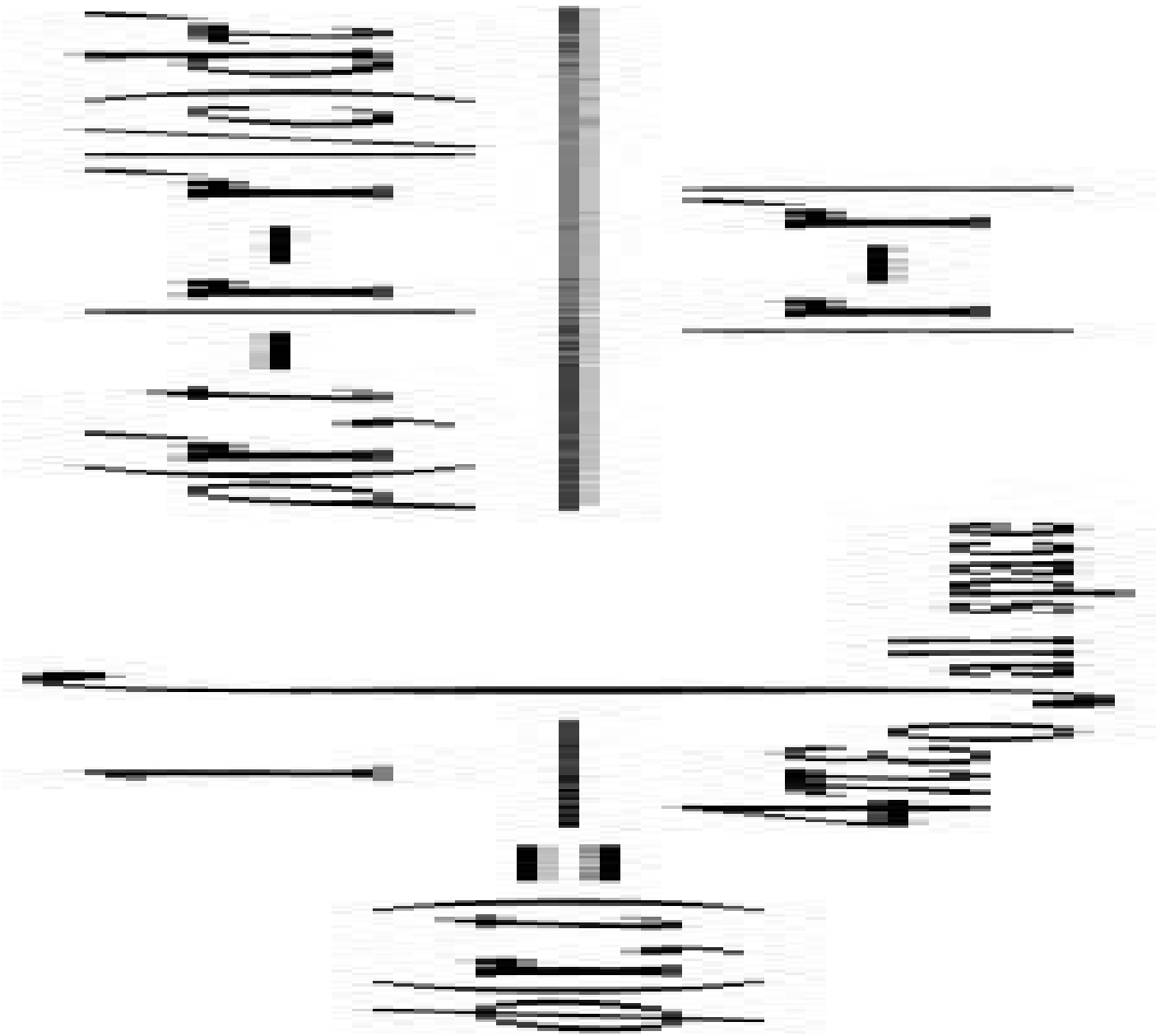
Substituting Equations (14.26), (14.27) and (14.28) into (14.24),

(14.29)

or, since $v\rho$ represents the current density j, which is pointing in the x-direction,

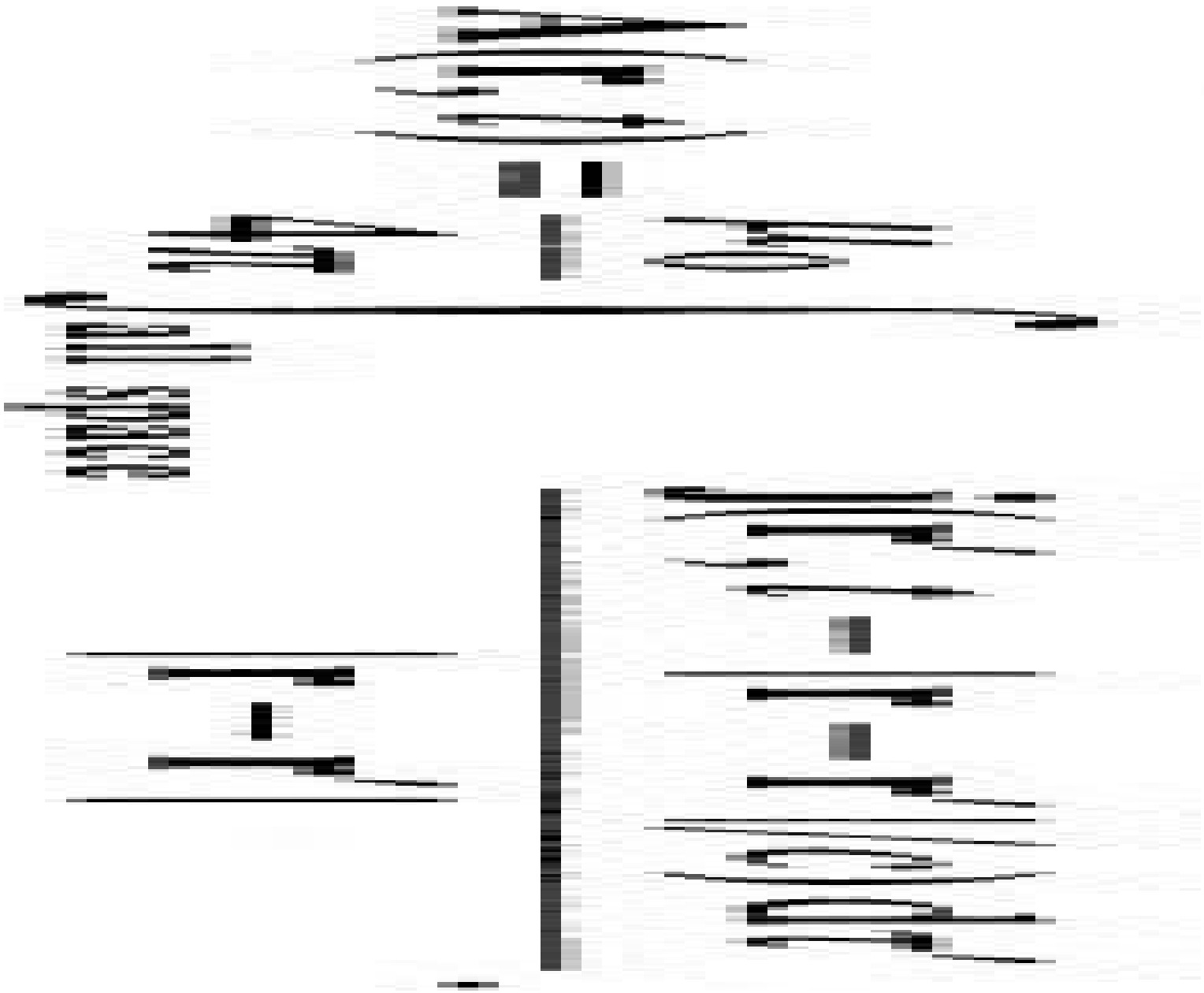
(14.30)

The potentials are the same as those due to a stationary charge $\rho d\tau$ and a steady current element $j\, d\tau$ situated at the origin at the earlier time $(t - r/c)$, and the volume element $d\tau$ can be taken as enclosing the origin, provided that ρ and j are evaluated at the time $(t - r/c)$. Equations (14.29) and (14.30) thus represent retarded potentials. All that remains to be done is to add together the contributions to ϕ and A from charge elements moving at different speeds in different directions, and to integrate over all of space. Returning to the notation of earlier chapters, we put r' for the position coordinates of a volume element with respect to a fixed origin and integrate Equations (14.29) and (14.30) to find



and

(14.31)



This is the same as the solution of Maxwell’s equations given in Equations (13.11) and (13.12).

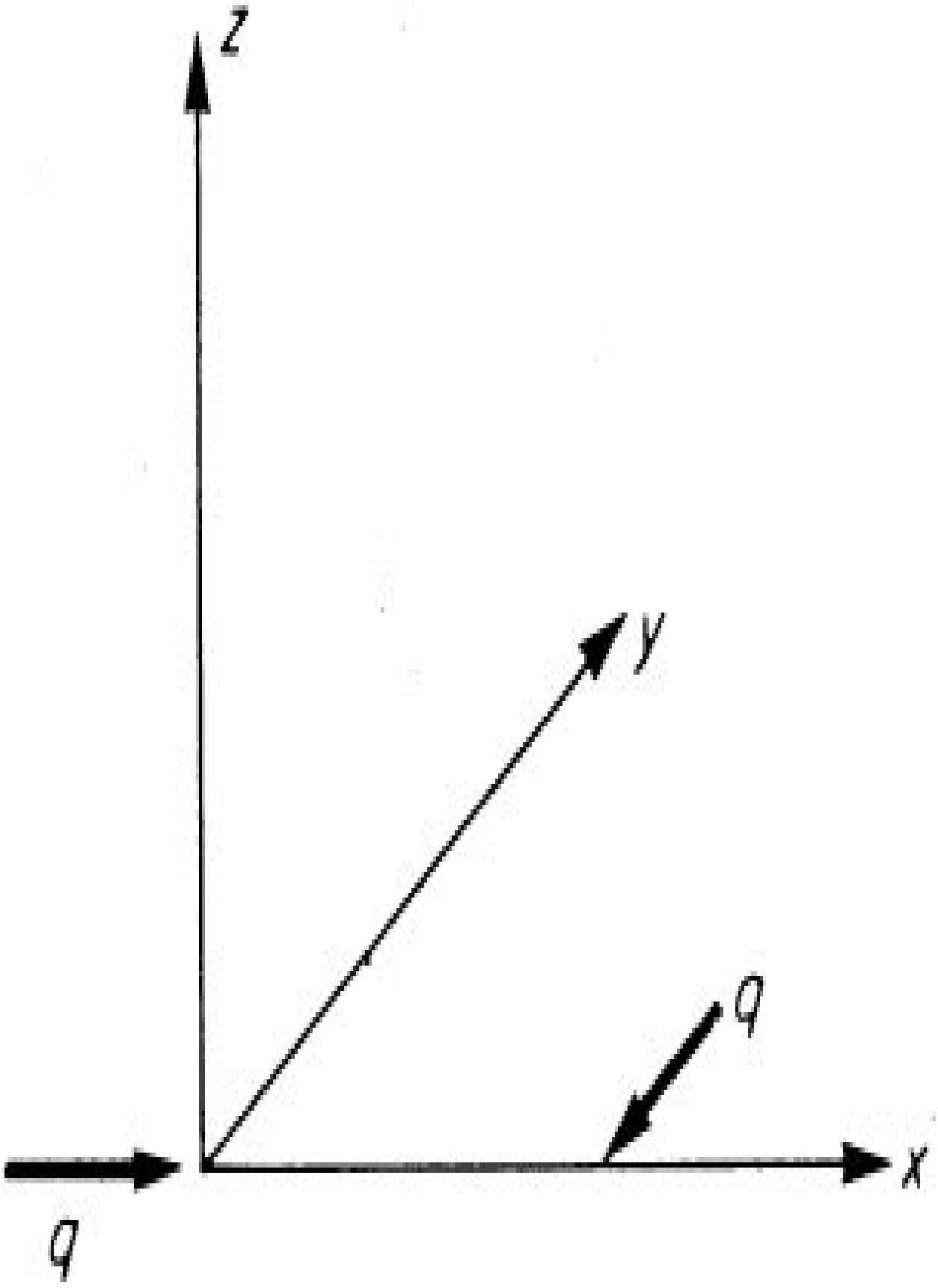
PROBLEMS 14

14.1 In its rest frame, a rectangular loop of wire with sides of length a and b is seen to be uncharged, and to carry a current I. Viewed from a frame moving at speed v in a direction parallel to the side of length a, what is the distribution of line charges on the loop? What is its electric dipole moment in this frame, expressed in terms of the magnetic dipole moment in the rest frame?

14.2 A cosmic ray proton of energy 10^{15} eV approaches the Earth moving perpendicularly to a magnetic field of magnitude $10-5$ T. What is the magnitude of the electric field experienced in the proton's rest frame?

14.3 Two point charges are moving in a frame S as shown in Figure 14.7. For an observer in S, the charge moving in the y-direction crosses the x-axis at the time when the other charge is at the origin. Using Equations (14.15) and (14.16) calculate the force he considers to be acting on the first charge. Is the force on the charge at the origin equal and opposite?

Figure 14.7.


z
y
q
q
x

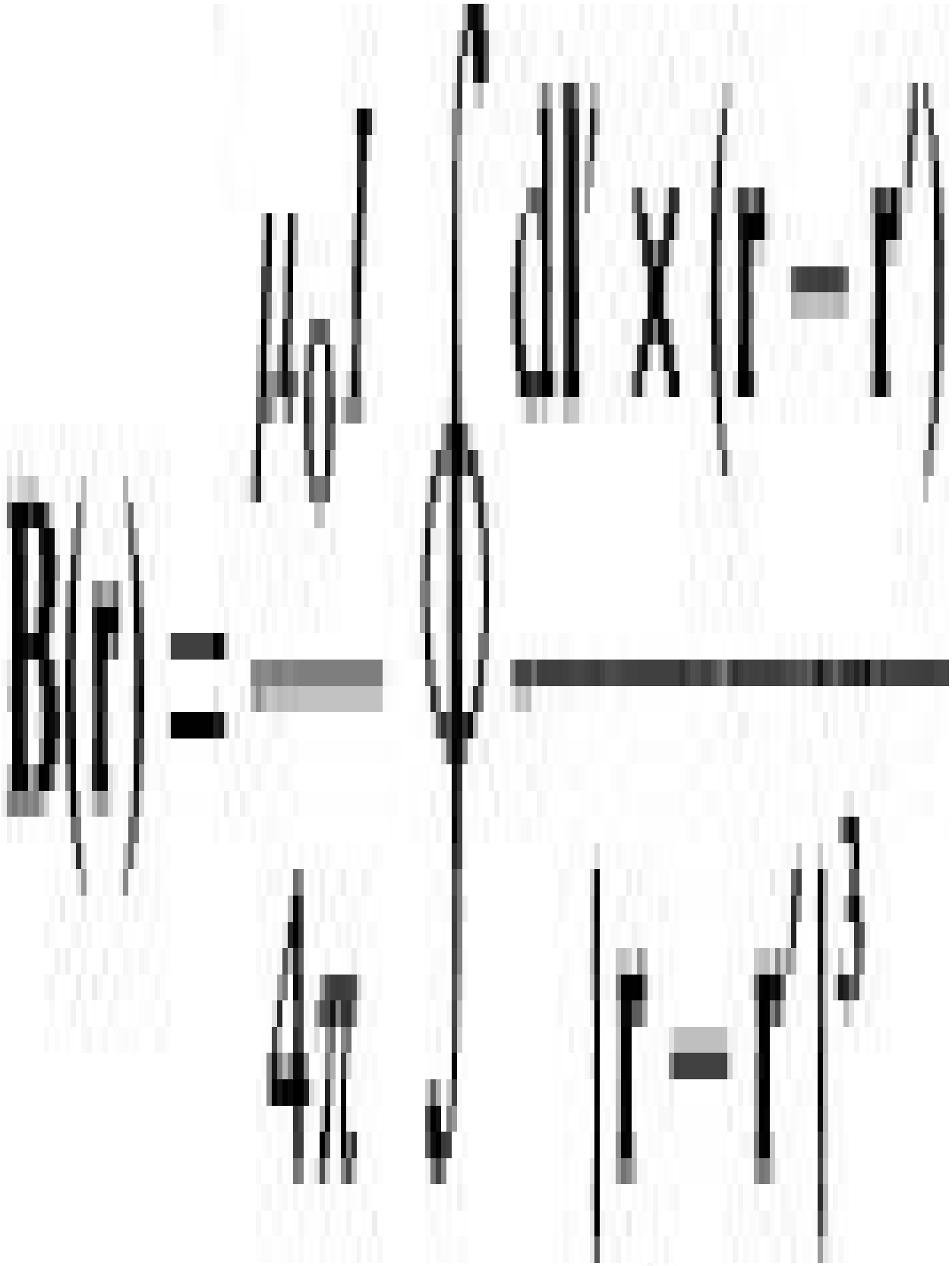
14.4 Generalize the result of Problem 14.3 to express the force between two circuits carrying steady current, and check that there is agreement with the Biot–Savart law.

* This experiment is described by F. J. M. Farley, J. Bailey and E. Picasso in Nature, 217, 17 (1968). The radiation was emitted by mesons moving at a speed greater than 0.99975c and the speed of the radiation was measured to be $(2.9979 \pm 0.0004) \times 10^8$ m/sec.

* It may seem paradoxical that the positive and negative charge densities on the wire transform differently, although the total charge on a body is invariant. But real steady currents always flow in closed circuits. The return half of the circuit (which may be too far away to make much difference to the field) has an opposite change in charge density, ensuring that the total charge on the complete circuit appears to be the same in both frames.

* The field is not strong enough to pull the electron straight away from the proton. But it is energetically favourable for the electron to separate from the proton in the direction opposite to that of the electric field. It does separate, because of the same quantum mechanical tunnelling effect which explains field emission (see Problem 1.12). Lorentz ionization provides us with another of the apparent paradoxes of relativity. Work must be done to ionize an atom, yet magnetic forces can do no work. However, although work is done by the electric field in the frame at rest relative to the atom, in the laboratory frame the kinetic energy alters; kinetic and potential energy are traded, but no external work is done.

* There is an inevitable confusion of notation because we are now using primes to distinguish moving frames of reference. The vector r of Equation (14.23) is the same as the r' of Equation (4.32) which is



Putting I dI′ = jdτ′, and evaluating this expression at the origin, this expression takes the same form as Equation (14.23).

APPENDIX A

Units

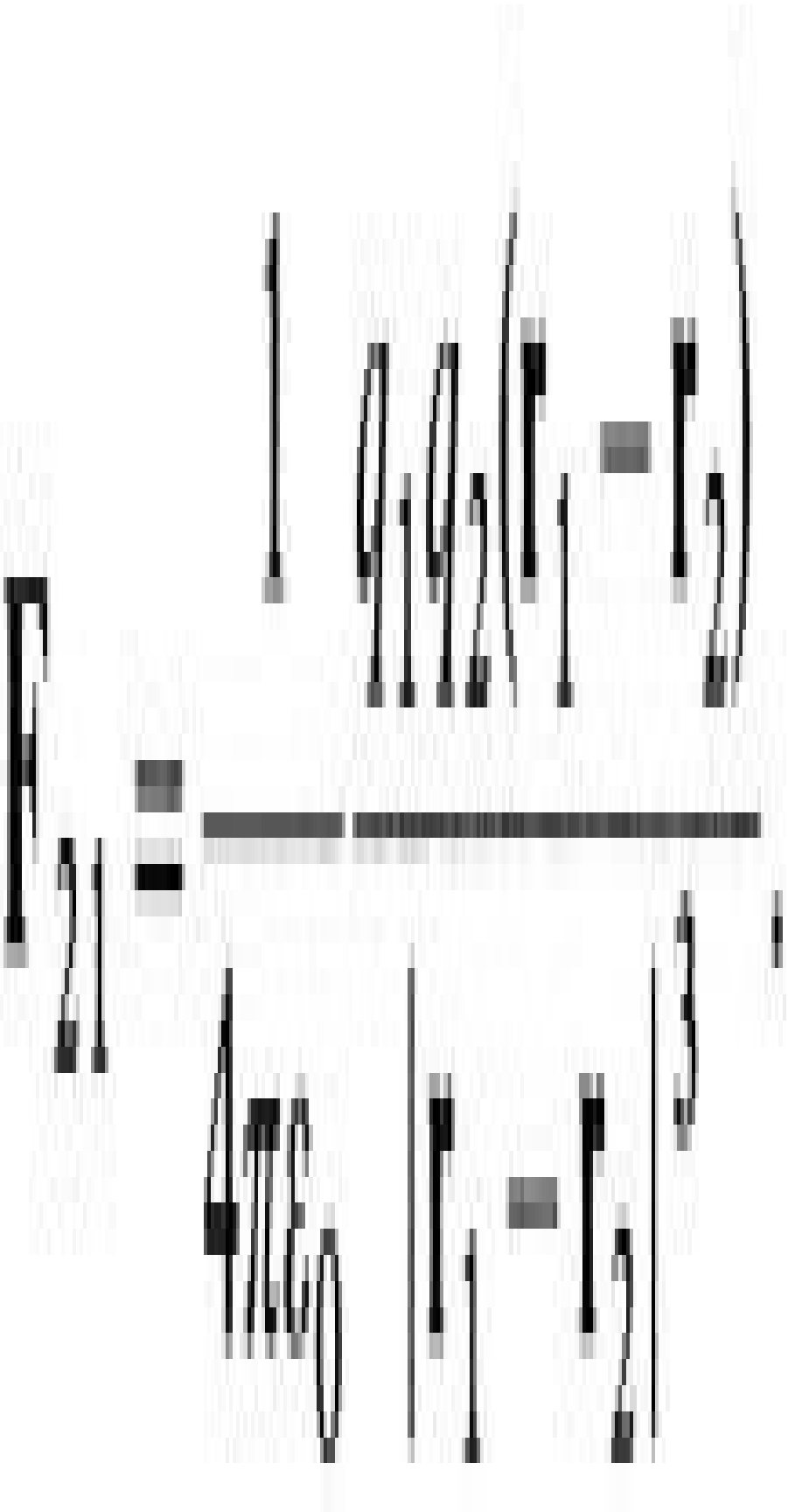
Throughout this book the units used for all physical quantities are the units of the Système International—SI units for short. SI units, which are recommended by the International Union of Pure and Applied Physics, are based on the metre, the kilogramme and the second as the fundamental units of length, mass and time. The SI electrical units are the same as the practical units in general use—amps, volts, ohms, etc. Section 1 of this appendix explains how the SI units are defined, and then describes the observations needed to set up precise electrical standards.

Other systems of units besides the Systéme International are still in common use. In particular, the Gaussian system of units is frequently used, especially in solid state physics. Gaussian units are defined in section A.2, and a summary is given of the change in form of the important equations of electromagnetism when they are expressed in Gaussian units. Finally, a table of conversion factors between the two sets of units is given in section A.3, together with a discussion of the dimensions of electrical units.

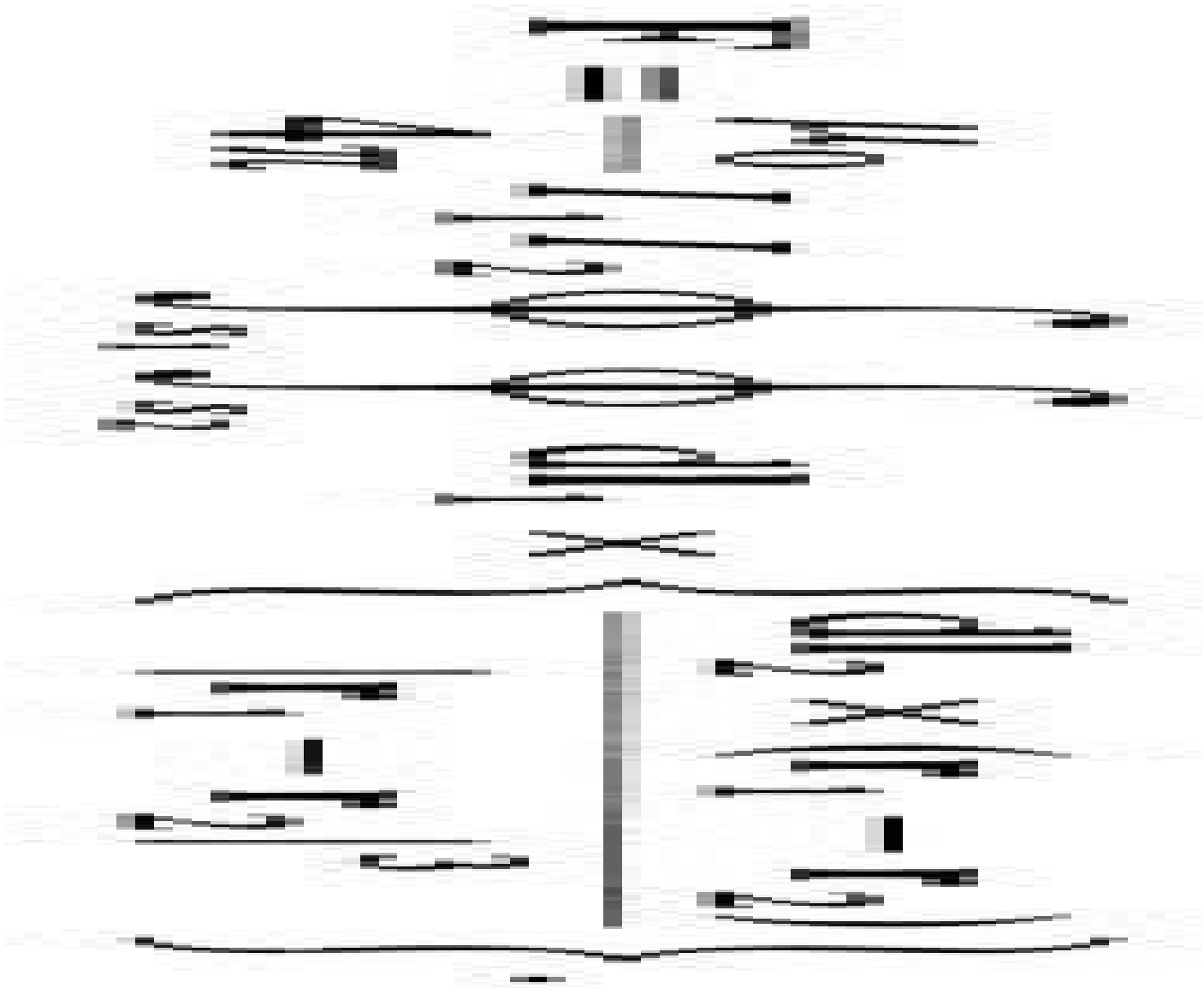
A.1 ELECTRICAL UNITS AND STANDARDS

A.1.1 The definition of the ampere

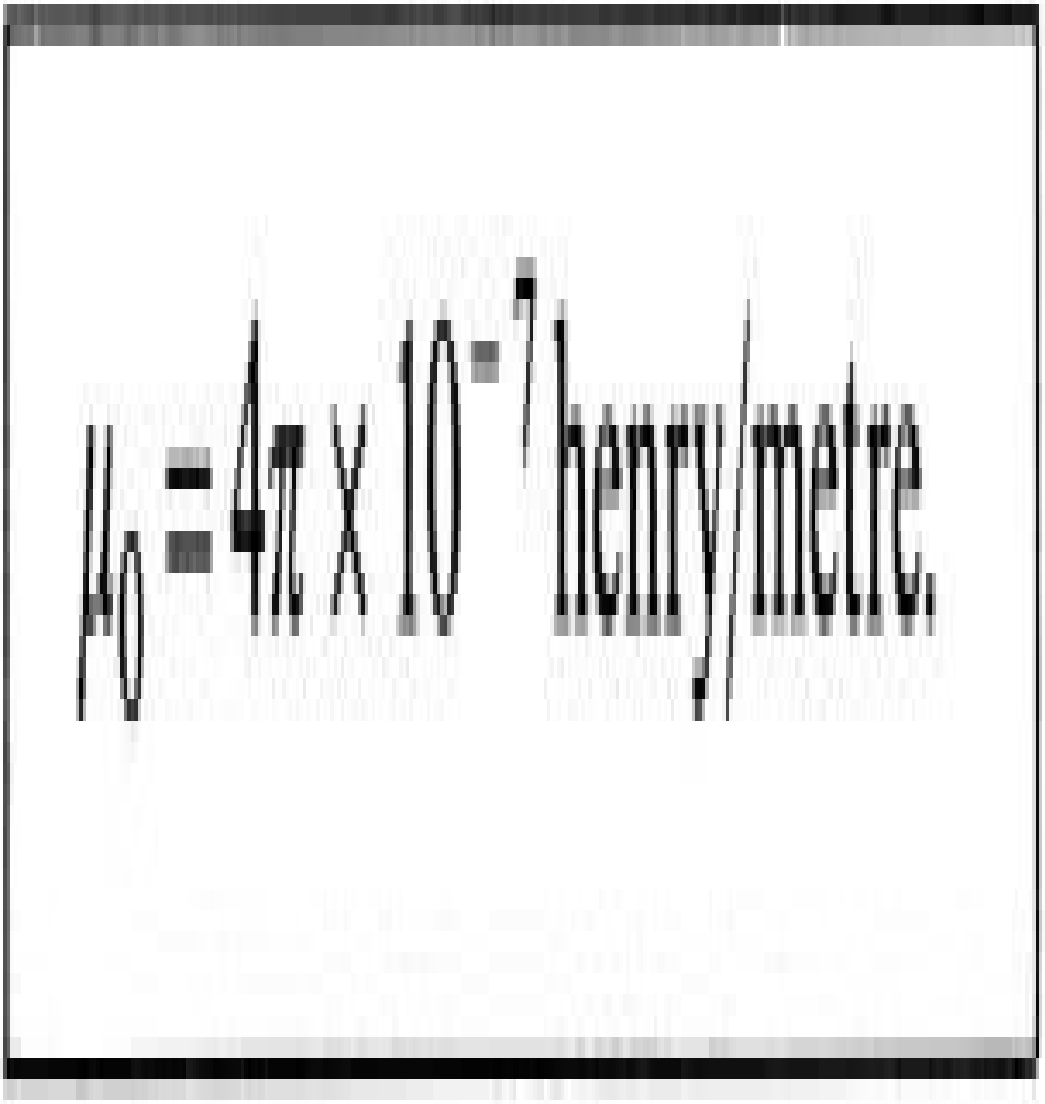
The constants ε0 and μ0 determine the magnitudes of SI electrical units in terms of the fundamental units of length, mass and time, The electrical units are directly linked to these fundamental units through the force laws for stationary charges and for steady currents. Coulomb's law (Equation (1.2)) gives the force between two stationary charges as



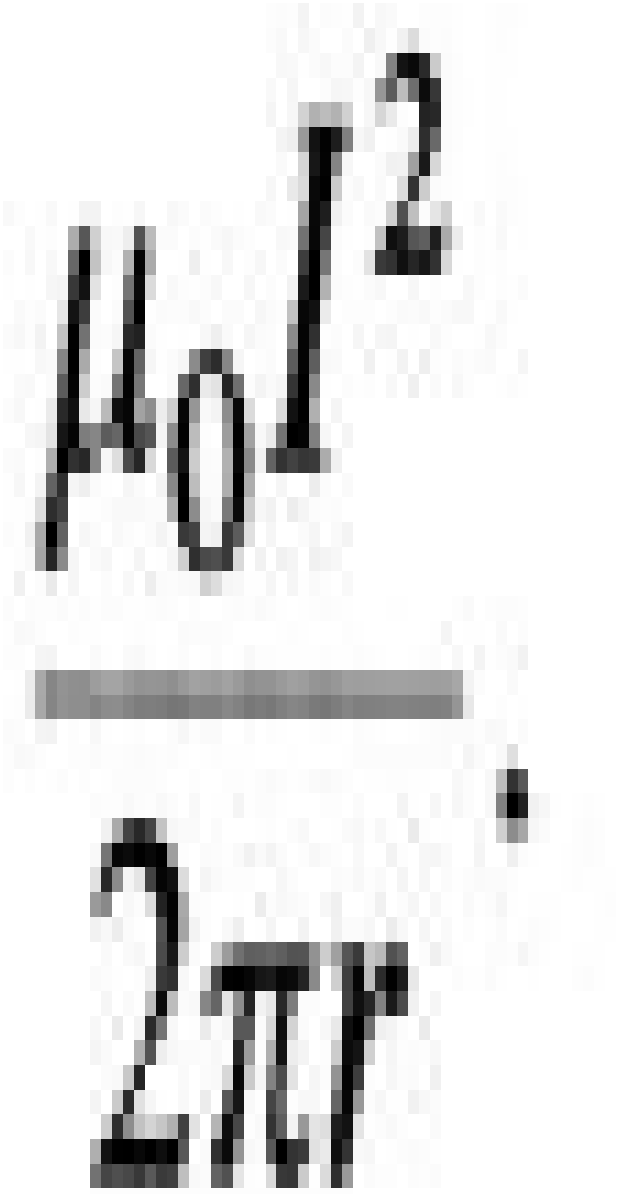
The force between circuits s1 and s2 made up of thin wires carrying steady currents I1 and I2 is found by combining the Biot-Savart law (Equation (4.32)) with Equation (4.50), which gives the force on a conductor placed in a magnetic field. The resulting expression is



The proportionality constants $1/4\pi\varepsilon_0$ and $\mu_0/4\pi$ occurring in these equations are not independent of one another, since the units of charge and current are not independent: the current flowing in a conductor is just the rate at which charge crosses a surface cutting the conductor. Consequently we are free to choose an arbitrary value for only one of the constants ε_0 and μ_0. In the Systéme International the permeability of free space μ_0 is given the value


$\mu_0 = 4\pi \times 10^{-7}$ henry/metre.

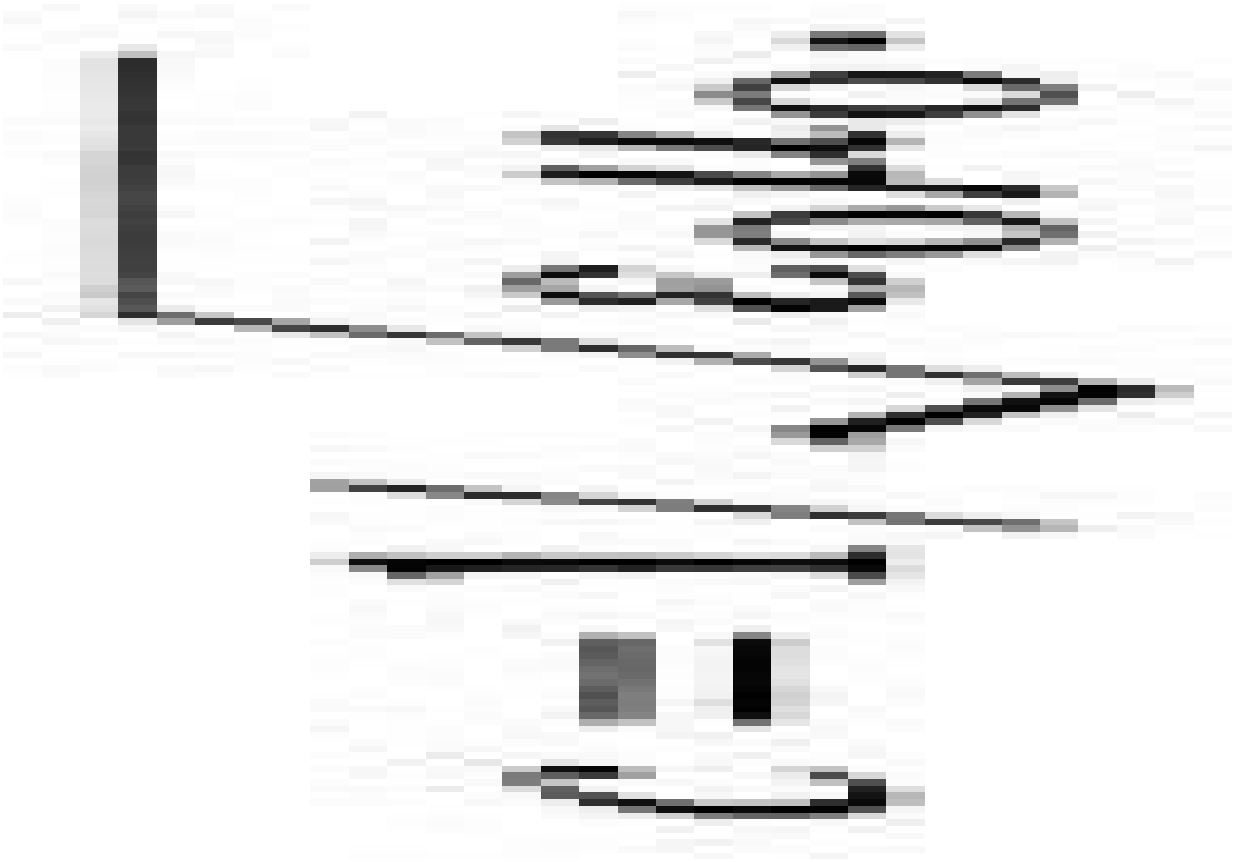
With this choice the unit of current defined by the Biot-Savart law is the ampere, which is already established as the practical unit. The definition can be given in physical terms by choosing circuits for which the integrals occurring in the Biot-Savart law can be carried out exactly. For example, if two thin parallel wires of infinite length are at a distance r apart and each carrying a current I, then the force per unit length between them is



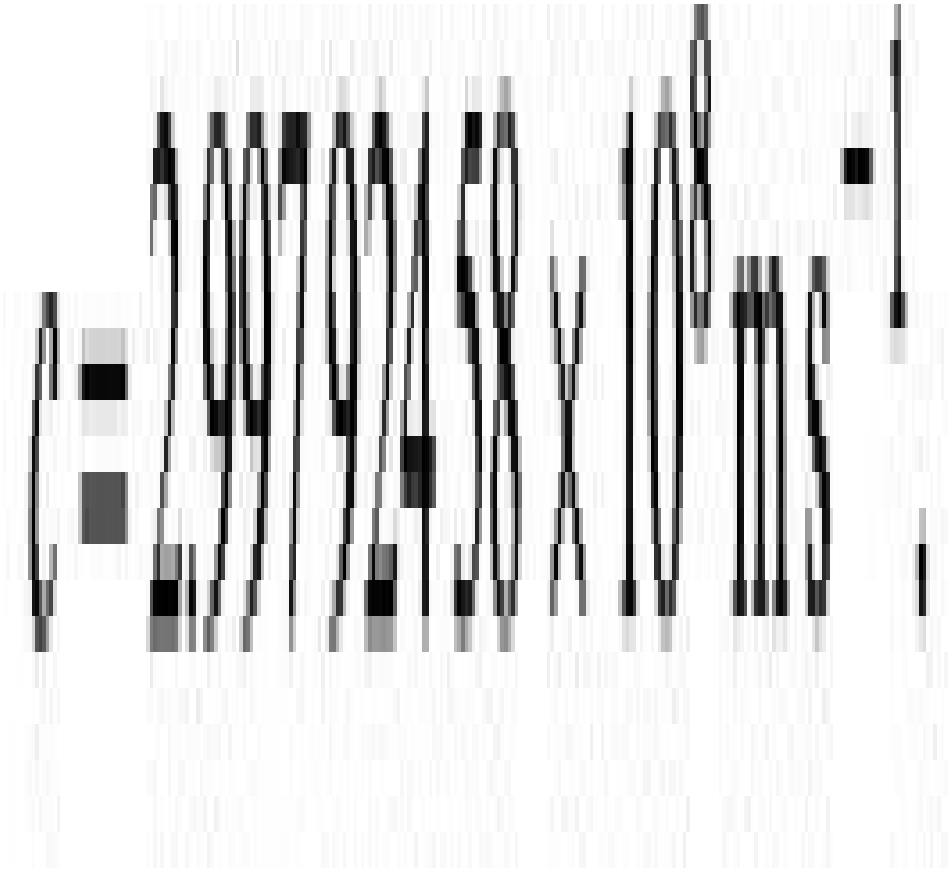
This result leads to the following physical definition of the ampere in terms of the SI units of force and length:

The ampere is the steady current which, if maintained in two straight parallel conductors of infinite length, of negligible circular cross-section, and placed 1 metre apart in vacuum, would produce between these conductors a force equal to 2×10^{-7} newtons per metre.

Once the ampere has been defined, the unit of charge, which is called the coulomb, must be such that one coulomb per second passes along a conductor if it is carrying a current of one ampere. Now that the unit of charge has been defined, the constant of proportionality in Coulomb's law is no longer arbitrary. It is shown in Chapter 11 that electromagnetic waves of any frequency—light waves, radio waves, etc.—propagate in free space with the same speed



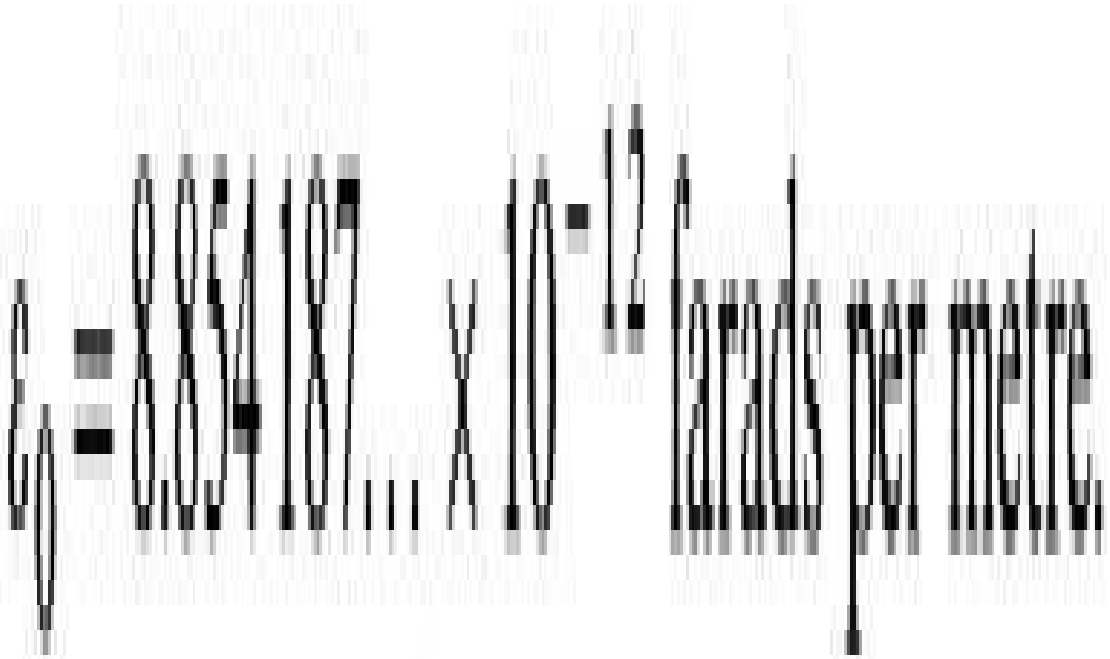
The units of length and time are not independent of one another. The second is defined by reference to caesium clocks, and is the duration of 9192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the caesium-133 atom. Since 1983 the unit of length, the metre, has been defined in terms of the distance light travels in one second. The speed of light is therefore not a measurable quantity, but is defined to have the exact value



This value is of course chosen to ensure that the unit of length is as close as possible to its old pre-1983 value.

Since μ_0 and c are both exact quantities, so is the permittivity of free space ε_0:

The numerical value of ε0, which can be given to any number of decimal places, is



A.1.2 Calibration and comparison of electrical standards.

The calibration of practical SI electrical units is done in two steps:

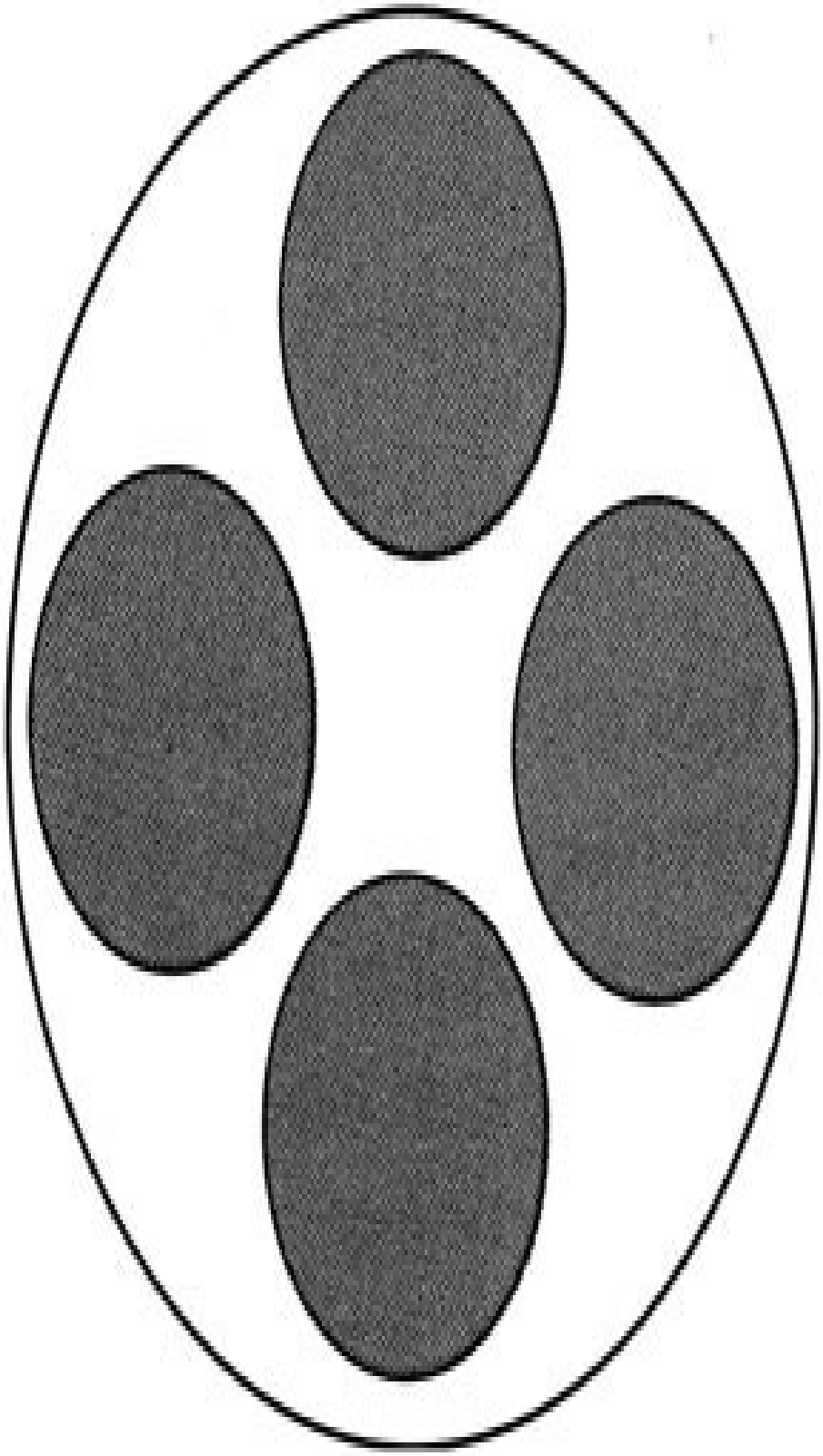
(i) absolute calibration, and

(ii) international agreement on the values of practical electrical standards which can be measured reproducibly and with great precision.

Absolute calibration

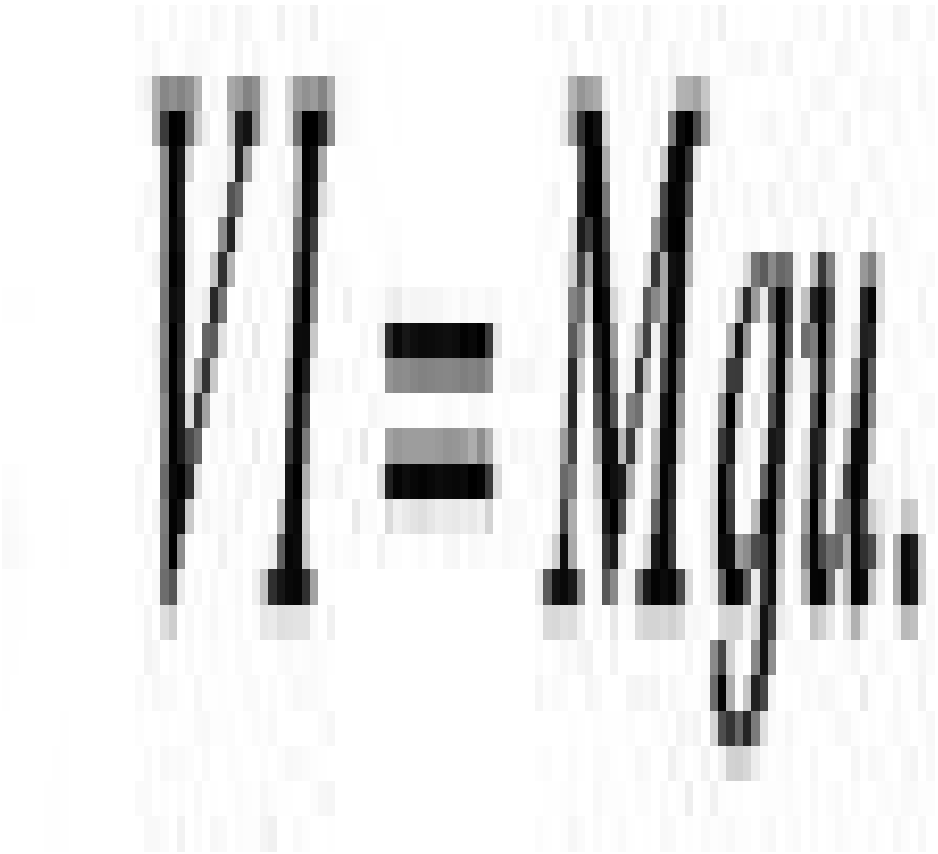
It is very difficult to make precise absolute measurements of electrical quantities in terms of the units of length, mass and time. Absolute measurements are normally carried out at standards laboratories such as the National Physical Laboratory in Great Britain or the National Institute for Standards and Technology in the United States. The force between two coils can be used to calibrate current, but because of the difficulties of measuring the geometry of the coils, this method cannot attain the required precision. An indirect method is now used for absolute calibration, depending on absolute measurements of resistance and of electrical power.

Figure A.1. Cross-section of a standard capacitor. The capacitance between opposite pairs of cylinders is measured.

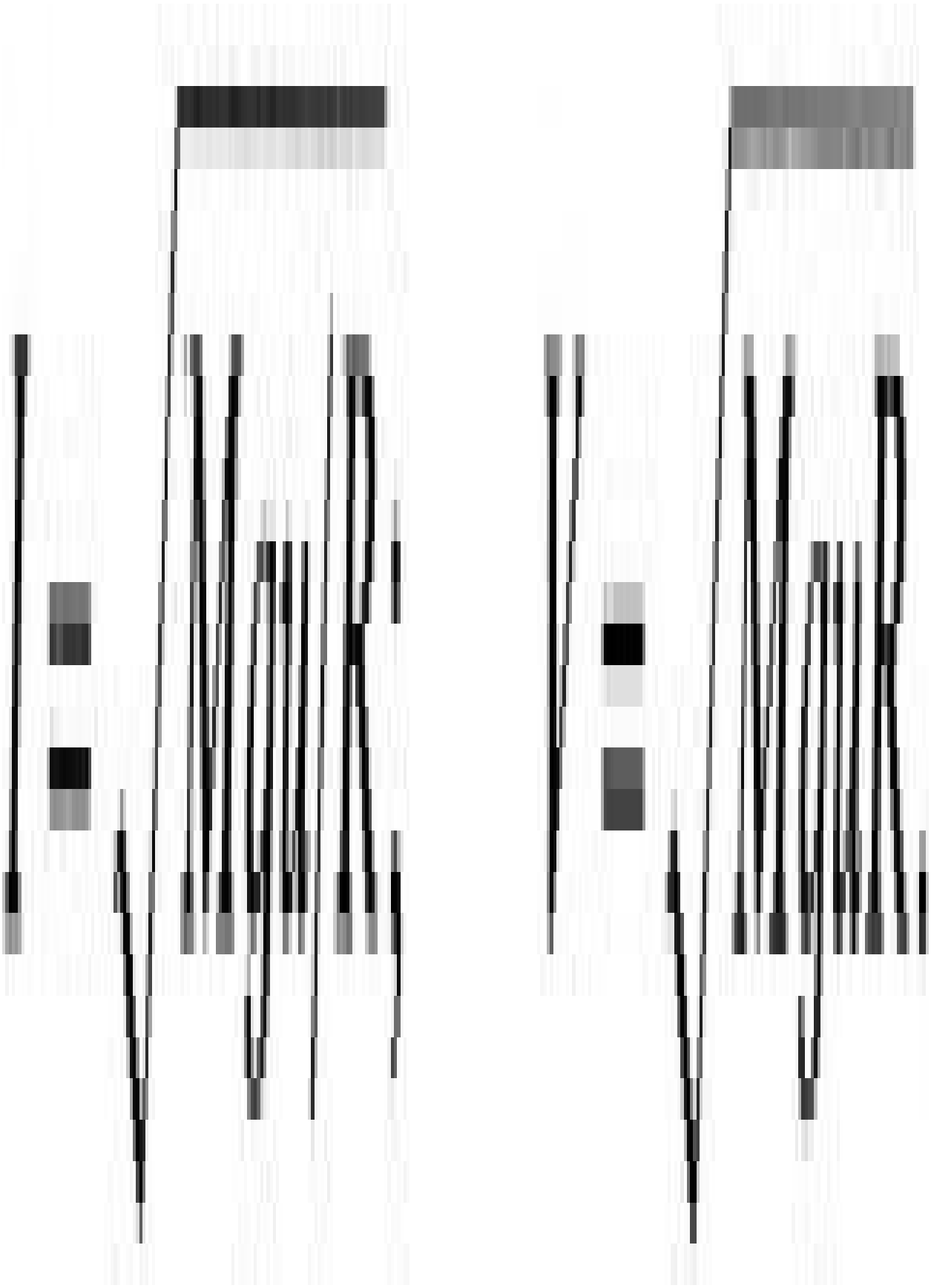


Absolute calibration of resistance is achieved by comparison with the impedance of a standard capacitor. Figure A.1 shows a cross-section of a standard capacitor which has four electrodes placed symmetrically around an axis through the centre of the figure. Provided that the electrodes are uniform, so that all cross-sections are the same, this capacitor has the remarkable property that, apart from end corrections, the capacitance per unit length between opposite pairs of electrodes is exactly $\varepsilon_0(\ln 2)/\pi$, no matter what shape or size the electrodes may be. Of course, in any real laboratory capacitor the end corrections are large, but they can be eliminated by making two measurements with different lengths of the capacitor. The capacitance of the standard capacitor is very small—only about 2pFm−1—and a number of capacitance comparisons and resistance comparisons must be made using bridge techniques to calibrate a practical standard resistor. However, the comparisons can be made very accurately, and the final precision is better than 1 part in 10^7.

The power calibration is made with coils suspended from the arm of a balance in a magnetic field. Two rectangular coils are mounted symmetrically so that each has one horizontal section in the strong field of a permanent magnet, and the other horizontal section outside the magnet in a weak field. When a current I flows through the coils, the resultant magnetic force is made to balance the gravitational force Mg on a calibrated mass M. The coils are then allowed to move through the field at a speed u, while carrying the same current I, and they generate a motional e.m.f. V. The equality of mechanical and electrical power requires that in SI units



The current I is also passed through a calibrated resistor and the voltage drop across the resistor is compared with the motional e.m.f. If V is equated to the voltage drop the current I would produce across a resistance R, then V = IR, and

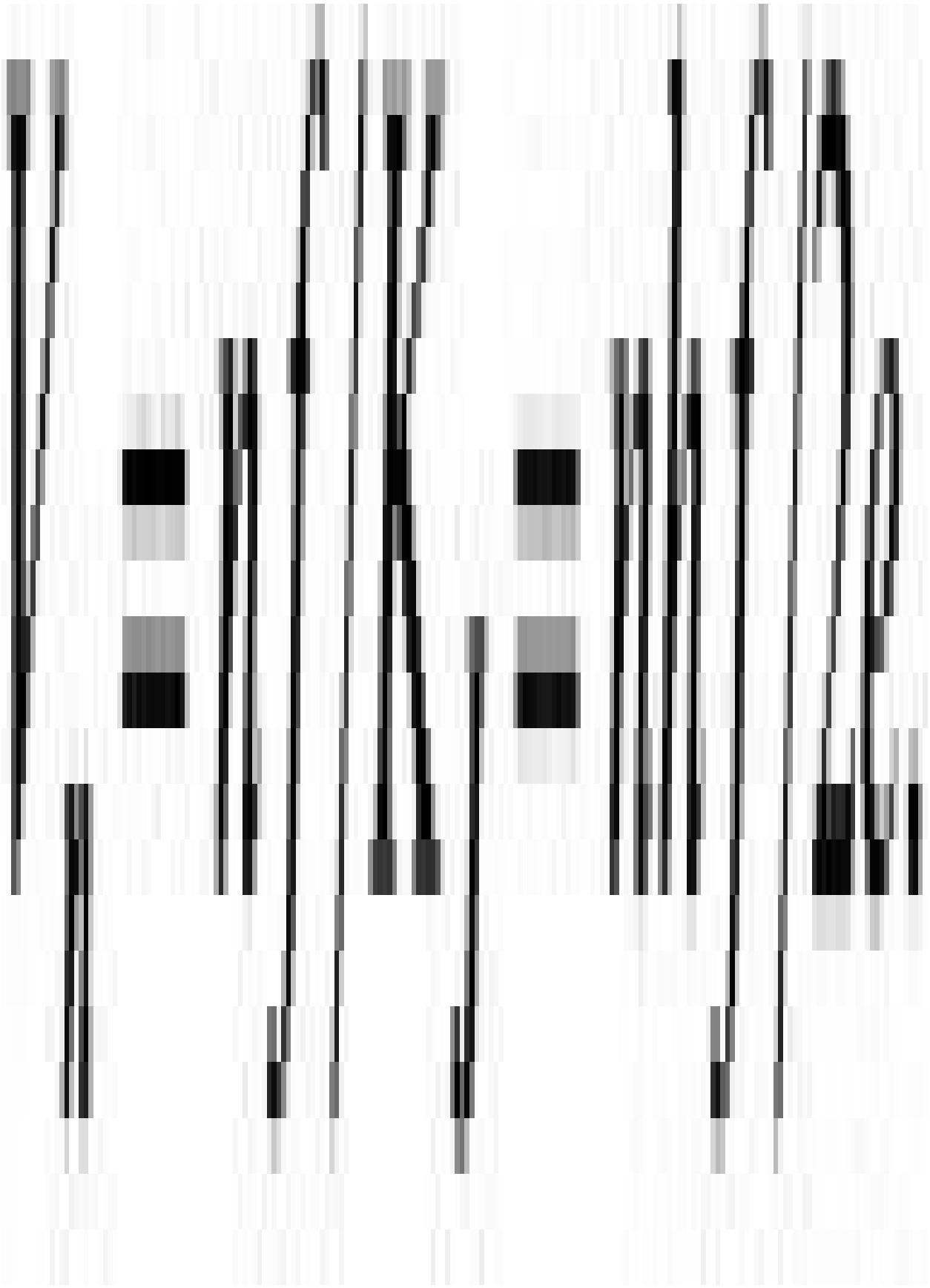


Both current and voltage are thus calibrated in terms of quantities which are known absolutely, without requiring to know anything about the dimensions of the coil or the strength of the magnetic field. By careful design and measurement this method of absolute calibration has achieved a precision of about 1 part in 10^7.

Practical standards

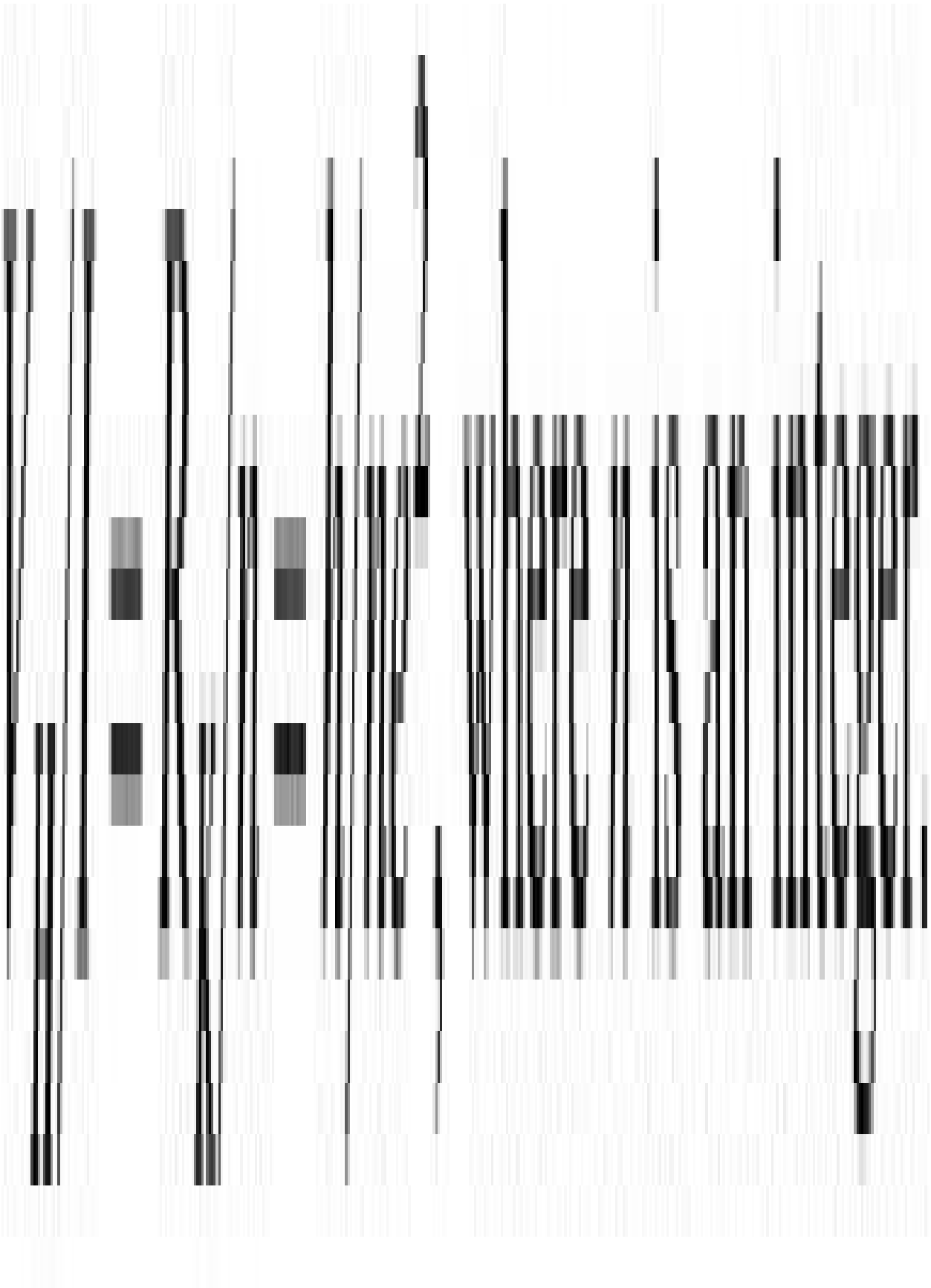
The absolute calibrations described above are extremely delicate and time-consuming, and are not suitable for the routine calibration of secondary electrical standards. Practical electrical standards are now defined with reference to two phenomena which depend only on the microscopic properties of atoms. They are the Josephson effect and the quantum Hall effect. Both are quantum effects, and it is not appropriate to discuss them in any detail in this book. We shall simply outline how they are used to establish electrical standards.

The Josephson effect occurs in superconducting devices called Josephson junctions which are exposed to microwave radiation of frequency f. The voltage across the junction does not vary smoothly with current, but exhibits steps at well-defined voltage intervals. The voltage Vn of the nth step is given by



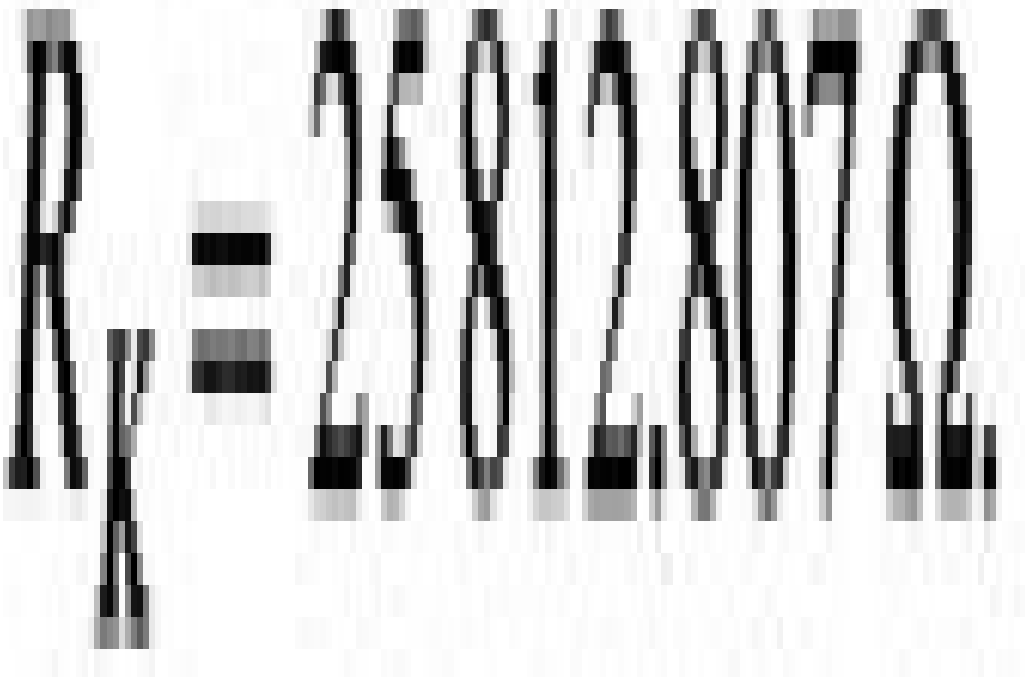
So far as is known, the Josephson constant KJ is exactly equal to h/2e, and the voltage steps can therefore be measured directly in terms of the frequency f.

The Hall effect has already been described in Chapter 4, where it was used to show how magnetic fields can be measured. The Hall voltage VH is related to the magnetic field B and to the current I flowing through the sample placed in the field. According to Equation (4.17), VH/I is proportional to B. This equation is normally correct, but in 1980 von Klitzing showed that in some circumstances semiconductor materials at very low temperature have Hall voltages which remain constant as the field is varied over a considerable range. These Hall voltage ‘plateaus’ occur where



Like the Josephson constant, the von Klitzing constant is determined by h and e, and both voltage and resistance are thus related to fundamental atomic constants.

The voltage steps in the Josephson effect and the plateaus in the quantum Hall effect are reproducible with a very high precision. They are suitable reference standards of voltage and resistance which can be set up in many different laboratories, and which have good long-term stability. At present, neither the accuracy of the absolute calibration of the SI electrical units nor the best values of e and h approach the precision of Josephson and quantum Hall measurements. Practical units of voltage and resistance have therefore been adopted by defining KJ and RK to be



These values are internationally agreed, and since 1990 they have been used to calibrate electrical quantities. No errors have been quoted for KJ and RK, since they have been assigned conventional values, used only for electrical measurements. True SI units are still defined by choosing $\mu 0 = 4\pi \times 10{-}7$, and in SI units Kj and RK have errors of a few parts in 10^7.

A.2 GAUSSIAN UNITS

The most commonly used system of electrical units other than the Systéme International is the Gaussian system (sometimes called the c.g.s. system). This system is based on the centimetre, the gramme and the second as the units of length, mass and time. Hence the unit of force is the dyne, and the unit of energy the erg. In the Gaussian system electrical units are defined by choosing the constant of proportionality in Coulomb's law to be unity. Coulomb's law therefore takes the form

(1.2G)

(In this section the equations are labelled with the same number as their first occurrence in the main text of the book, but with the suffix G to indicate that they have been expressed in Gaussian units.) The unit of charge is called the e.s.u. or electrostatic unit. From Coulomb's law, when two charges each of magnitude 1 e.s.u. are separated by a distance of 1 cm, each experiences a force of 1 dyne.

Electrostatics

Most of the equations of electrostatics have a different form in Gaussian units because of the removal of the factor $1/4\pi\varepsilon 0$ from Coulomb's law. The important electrostatic equations are listed below in their Gaussian form, with comments where necessary on the Gaussian units for various quantities. Gauss' law becomes

(1.20G)

The electric field E is measured in statvolts/cm, and is related to the potential (measured in statvolts) by the equation

(1.26G)

The potential energy of an assembly of point charges is

(1.36G)

The energy density of an electric field E is $E^2/8\pi$, leading to a total energy

(1.42G)

The dipole moment per unit volume P has the same dimensions as the field E, and in an isotropic material of relative permittivity ε has the value

(2.5G)

The electric susceptibility is now defined by $P = \chi_e E$, so that

(2.7G)

(This is especially confusing: note that the dimensionless quantity χ_e has different values in the two systems of units.) The electric displacement

(2.20G)

satisfies Gauss's law for the free charge density;

(2.19G)

The total electrostatic energy stored in a volume V is

(2.26G)

Finally, Poisson's equation becomes

(3.2G)

for a material with uniform relative permittivity ε.

Currents and magnetic fields

In the Gaussian system of units the Lorentz force is written in such a way that the magnetic field B has the same dimensions as the electric field E, namely force per unit charge. The force law for a charge q moving in an electromagnetic

field is

(4.16G)

where c is the speed of light. The speed of light also occurs in the Biot–Savart law:

(4.32G)

In magnetic material with a dipole moment per unit volume M, the magnetic intensity H is defined as

(5.27G)

The relative permeability μ is defined by

(5.32G)

and the magnetic susceptibility χ_H by

(5.31G)

Notice that as for the electric susceptibility, the magnetic susceptibilities in SI and Gaussian units differ by a factor 4π. Ampère's law for a steady density of free current j_f takes the form

(5.30G)

The total energy stored in a volume V by a steady magnetic field is

(6.35G)

The chief practical objection to the Gaussian system is that its units for the quantities required in circuit analysis do not coincide with the universally accepted units based on volts and amps. The statvolt is rather large (about 300 volts) and the e.s.u./sec is a very small unit of current (about $3.3 \times 10–10$amps).

The units of impedance are based on the usual equations:

Resistance R: $V = IR$

Capacitance C: $Q = CV$

Inductance L: $V = L\frac{dI}{dt}$.

The units of R, C and L have no special names, and are referred to as ‘Gaussian (or c.g.s.) units of resistance’ etc.

Maxwell’s equations

Maxwell’s equations in Gaussian units take the form

(10.10G)

(10.11G)

(10.12G)

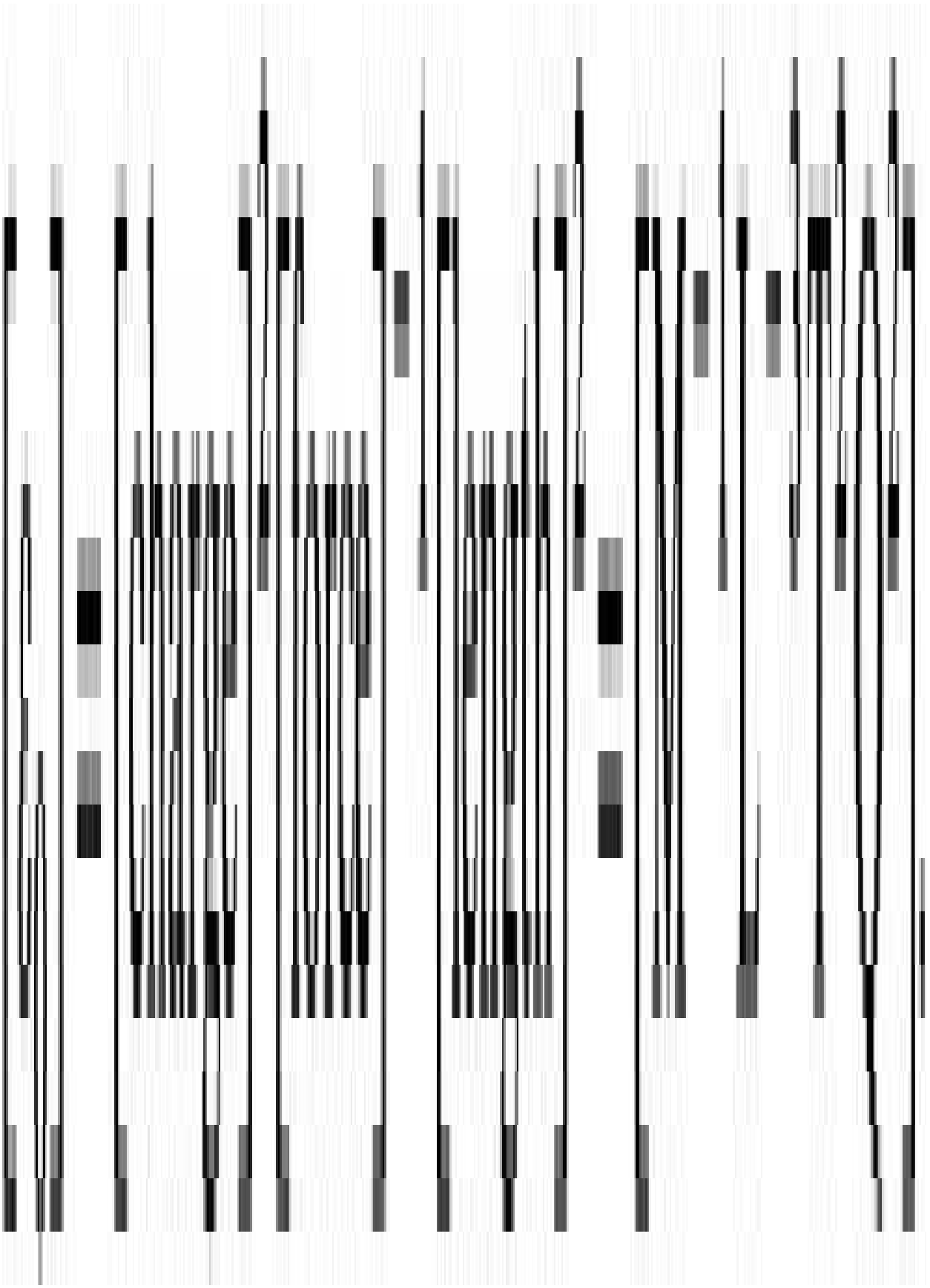
(10.13G)

A.3 CONVERSION BETWEEN SI AND GAUSSIAN UNITS

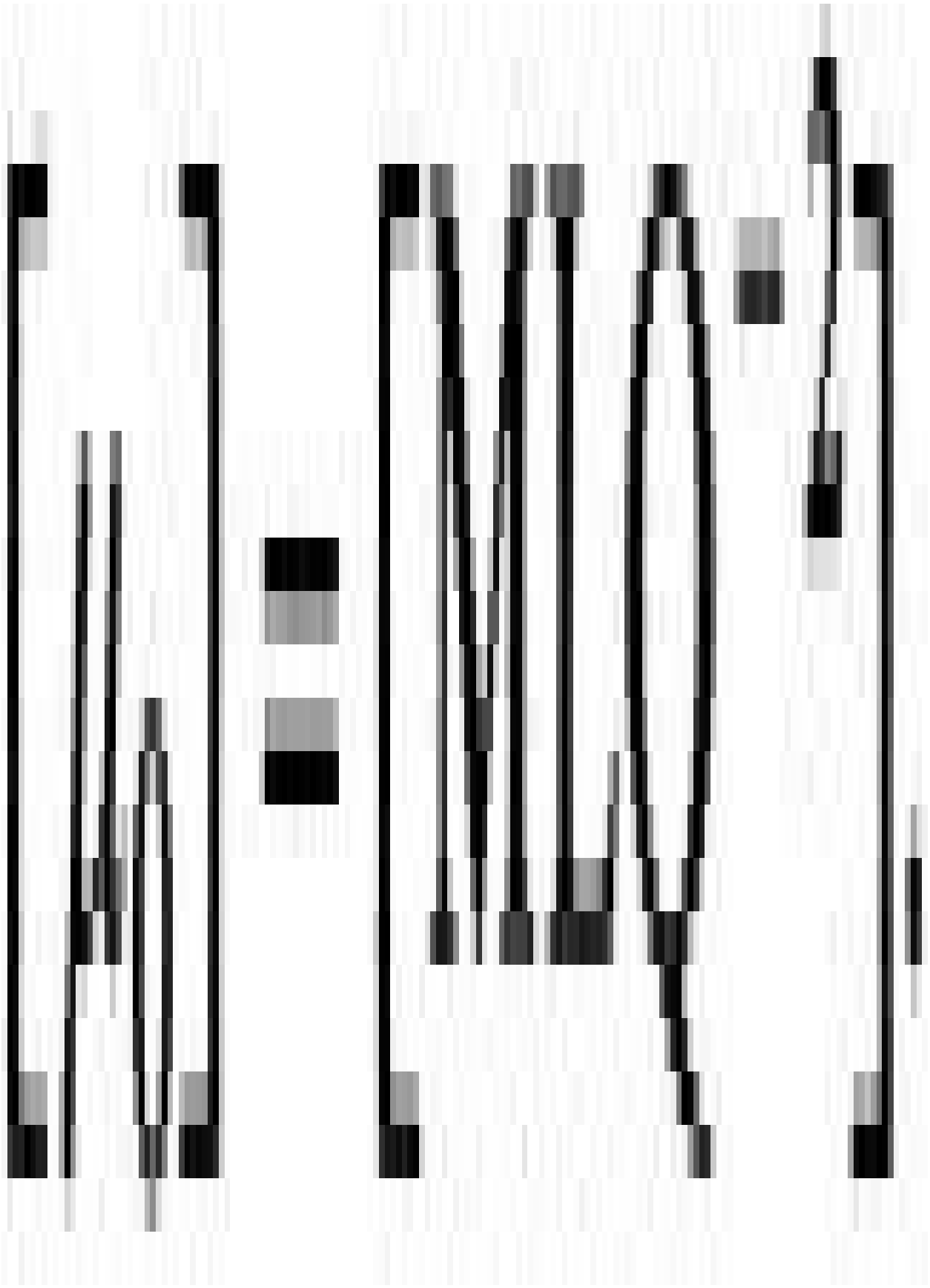
A table is given overleaf of the conversion factors between the two systems of units.

Dimensions

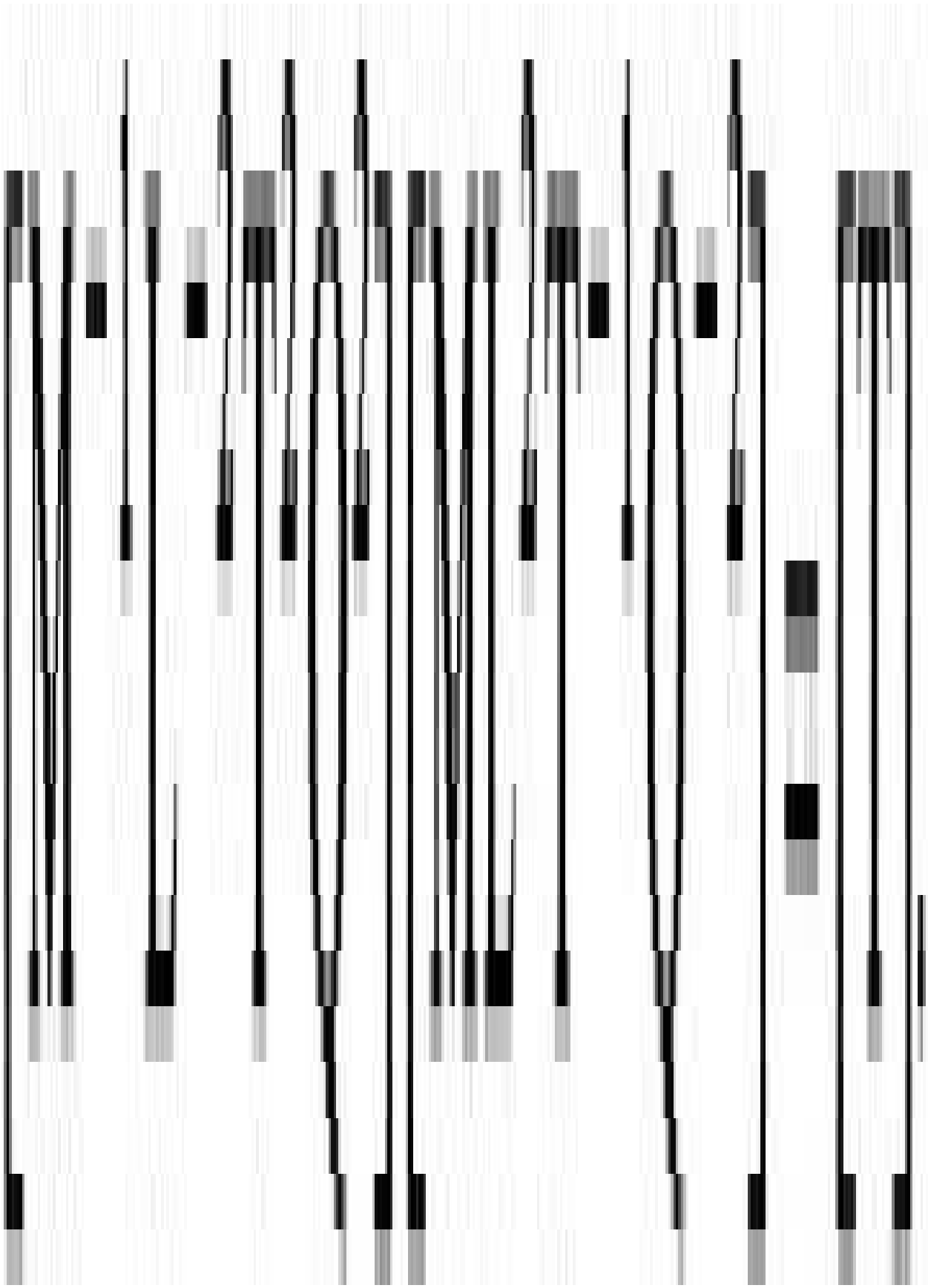
Charge has been given the dimension [Q] in the table. Coulomb's law then requires that the permittivity of free space has the dimensions



Since $\varepsilon_0\mu_0 = 1/c^2$, the permeability of free space has the dimensions



Dimensional analysis is applied to the electrical quantities in the usual way. Thus the product CR has dimensions



Notice that the dimensions in the third column do not apply to Gaussian units, since the two sets of units are used in electromagnetic equations of a different form. In Gaussian units the fields D and E have the same dimensions, as do the fields B and H. However, since $B = \mu H$ in Gaussian units, a field of one gauss corresponds to a field of one oersted only in the absence of magnetic materials.

CONVERSION FACTORS BETWEEN SI AND GAUSSIAN UNITS

Quantity	SI unit	Dimensions of SI unit	Gaussian unit	No. of Gaussian units in one SI unit
Force	newton	$[MLT^{-2}]$	dyne	10^5 (i.e. 10^5 dyne = 1 newton)
Energy	joule	$[ML^2T^{-2}]$	erg	10^7
Power	watt	$[ML^2T^{-3}]$	erg/s	10^7
Charge	coulomb	$[Q]$	e.s.u.	3×10^9
Current	ampere	$[T^{-1}Q]$	e.s.u./s	3×10^9
Potential	volt	$[ML^2T^{-2}Q^{-1}]$	statvolt	1/300
Capacitance	farad	$[M^{-1}L^{-2}T^2Q^2]$		9×10^{11}
Resistance	ohm	$[ML^2T^{-1}Q^{-2}]$		$1/(9 \times 10^{11})$
Inductance	henry	$[ML^2Q^{-2}]$		$1/(9 \times 10^{11})$
E	volt/metre	$[MLT^{-2}Q^{-1}]$	statvolt/cm	$1/(3 \times 10^4)$
D	coulomb/sq. metre	$[L^{-2}Q]$	statvolt/cm	$12\pi \times 10^5$
B	tesla	$[MT^{-1}Q^{-1}]$	gauss	10^4
H	ampere/metre	$[L^{-1}T^{-1}Q]$	oersted	$4\pi \times 10^{-3}$

In this table c has been taken as 3×10^8 m/sec.

APPENDIX B

Fields and differential operators

A scalar or vector function of position is called a field. Fields may be operated on by differential operators, called div, grad and curl, which generate new fields. The operators div, grad and curl are defined in section B.l. Formulae are given in section B.2 for the fields generated by these operators, expressed in different coordinate systems. Finally, some useful identities are listed in section B.3. The formulae in this appendix are given with little explanation: it is intended as a reminder of the properties of fields and operators, not as a substitute for a mathematical text.

B.1 THE OPERATORS DIV, GRAD AND CURL

Divergence

The divergence operator acts on a vector field, and the resultant is a scalar field. The divergence of an arbitrary vector field F(r), written div F or ∇·F, represents the outward flux of F per unit volume at the point r.

The flux of F out of a closed surface S enclosing a small volume δτ is ∫5F·dS, taking dS to be the outward normal to the element of area dS. In the limit, as δτ tends to zero, the outward flux of F per unit volume is independent of the shape of S, provided that F is smoothly varying. Thus

(B1)

As an example, consider the divergence of the current density j(r). Current density is the rate of flow of charge per unit area, and ∫sj dS is therefore the net rate of removal of charge from the volume within S. It follows that div j is zero in any conductor which remains electrically neutral throughout its volume, whatever the pattern of currents flowing in it.

Gradient

The gradient operator acts on a scalar field, and the resultant is a vector field.

The gradient of an arbitrary scalar field $\psi(r)$ is written as grad ψ or $\nabla\psi$. At any point, grad ψ lies along the direction in which ψ is most rapidly increasing, and its magnitude is equal to the slope of ψ in that direction.

If e is a unit vector (Figure B.1) in the direction of the line PQ joining two points P and Q with position vectors r and $(r + \delta r)$, then at P the component of grad ψ along e is

(B2)

For example, the electrostatic potential $\phi(r)$ is a scalar field, and the electrostatic field $E = -$ grad ϕ is the force per unit charge pushing positive charges towards regions of lower potential energy.

Curl

The curl operator acts on a vector field, and the resultant is another vector field. The curl of an arbitrary vector field F(r) is written as curl F or $\nabla \times F$. The curl of a vector field is a measure of its vorticity. If the field lines of a vector field circulate around a central line like the streamlines of a whirlpool around its vortex, the curl of the vector lies in the direction of the central line.

If e is a unit vector perpendicular to a small surface of area δS located at r, then

Figure B.l.

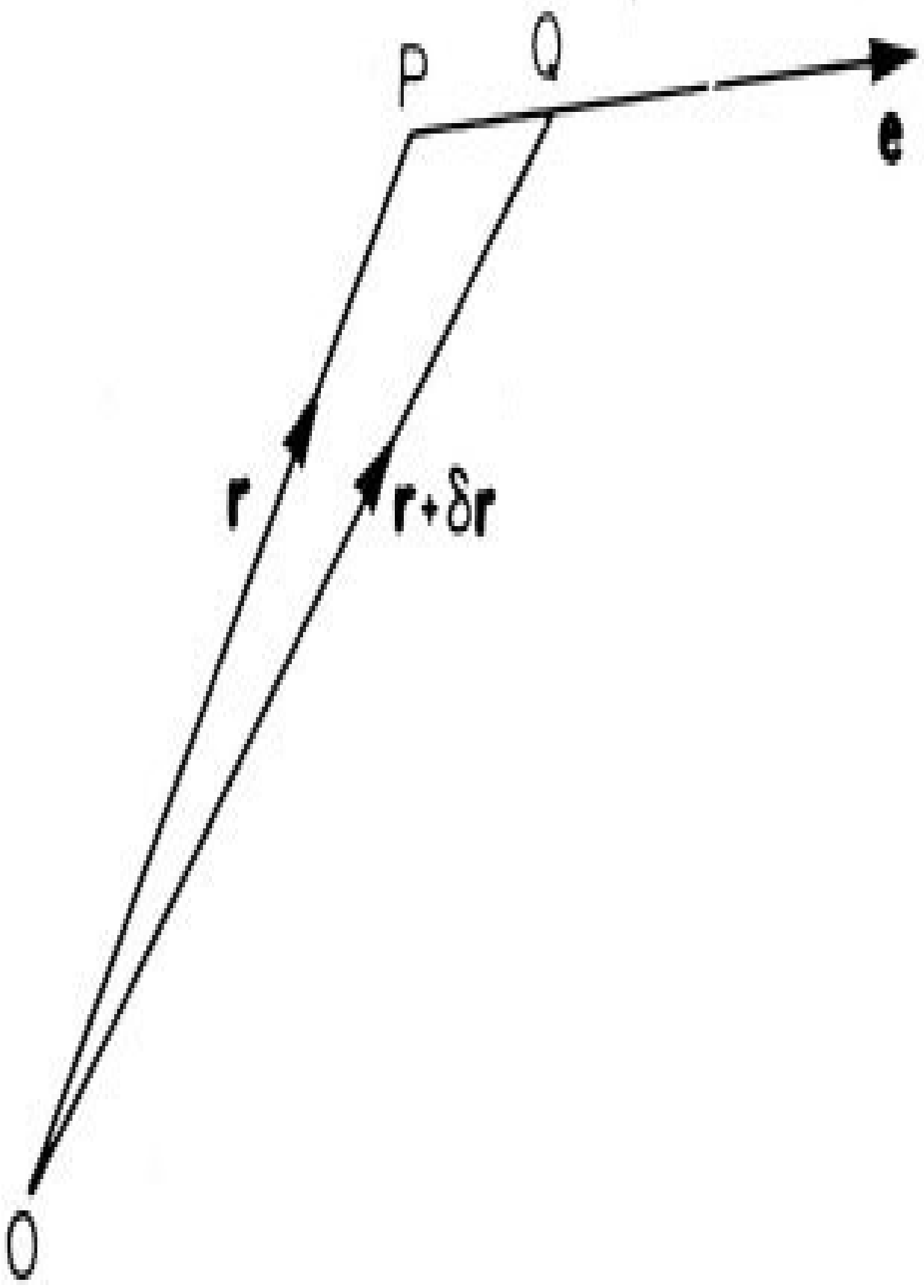
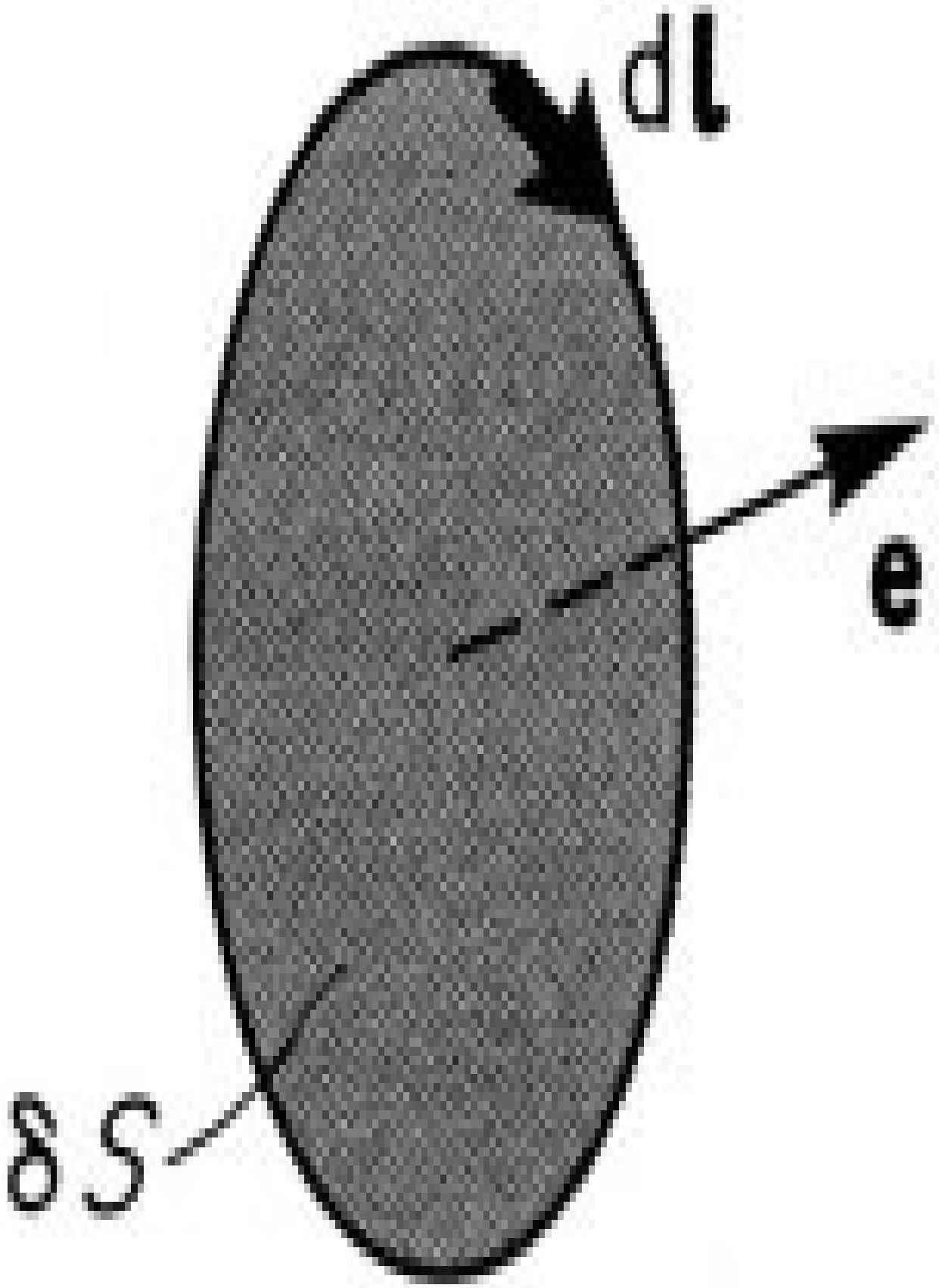

P
Q
e
r
r+δr
O

Figure B.2.


dl
e
δS

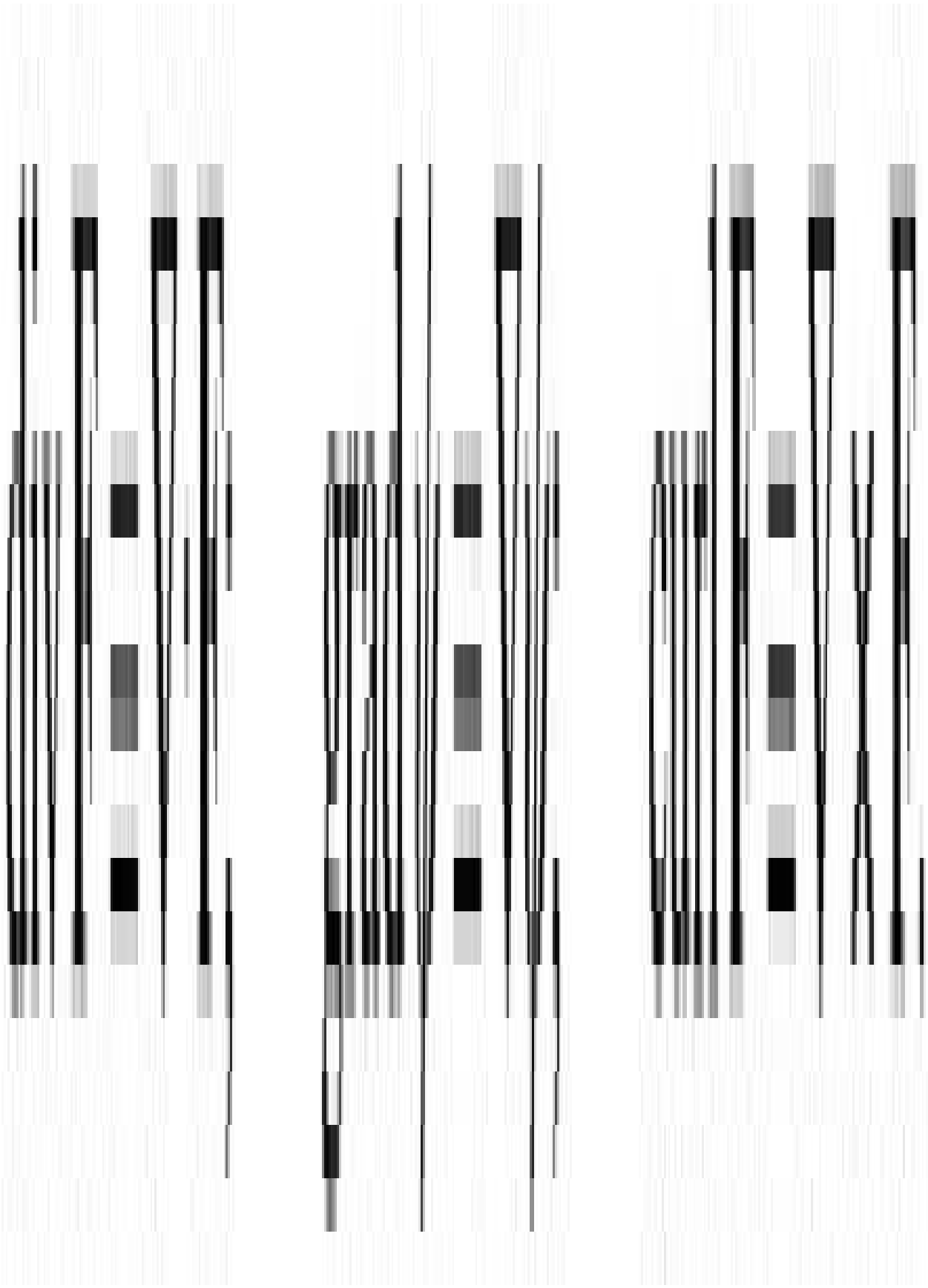
the component of curl F along e is

(B3)

where the line integral is evaluated around the perimeter s of the area δS, in a right handed sense with respect to e as indicated in Figure B.2. An example of a field which can have non-zero curl is the magnetic field B. The field lines of B close up around the currents which generate the field. But in regions where there is no free current or magnetization current curl B is zero.

The fields div F, grad ψ and curl F can be expressed in a concise way by using the vector differential operator ∇. The symbol ∇ is pronounced del. Using the ∇ notation, div, grad and curl are written equivalently as

(B4)



Remembering that ∇ is a vector quantity serves to remind us that $\nabla \cdot F$ (a scalar product) is a scalar field, whereas $\nabla\psi$ and $\nabla \times F$ are vector fields. (The symbol '≡' means 'is equivalent to'.)

The Laplacian operator ∇^2

By operating on a scalar field $\psi(r)$ with the gradient and divergence operators in turn, another scalar field div grad $\psi \equiv \nabla \cdot (\nabla\psi) \equiv \nabla^2\psi$ is formed. Since each of the three components of a vector field F are independent, the Laplacian operator ∇^2 may be allowed to operate separately on the three components, to form the components of a vector field $\nabla^2 F$.

B.2 FORMULAE IN DIFFERENT COORDINATE SYSTEMS

The fields resulting from operating with differential operators must be worked out in a coordinate system suited to each particular problem. Formulae for the fields are given below in Cartesian, spherical polar and cylindrical polar coordinates.

Cartesian coordinates

In Cartesian coordinates, a vector field F(x, y, z) is resolved at any point into its components Fx, Fy and Fz along the directions of the x-, y- and z-axes. In Cartesian coordinates the operator ∇ has components $\partial/\partial x$, $\partial/\partial y$ and $\partial/\partial z$. Expressing ∇ in terms of the unit vectors ex, eyand ez,

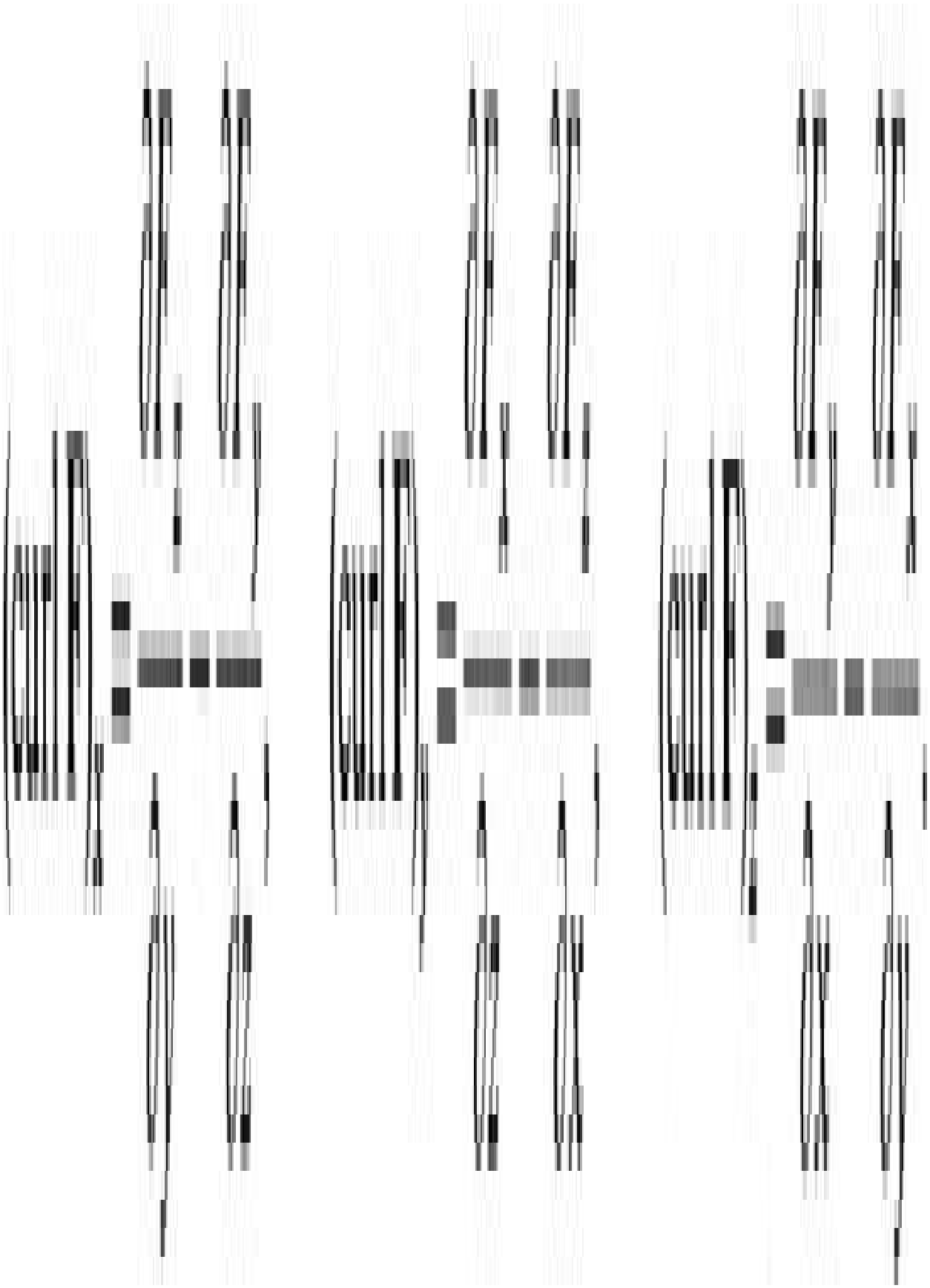
(B4)

The usual rules for forming a scalar product may be applied to ∇ in Cartesian coordinates, and we find

(B5)

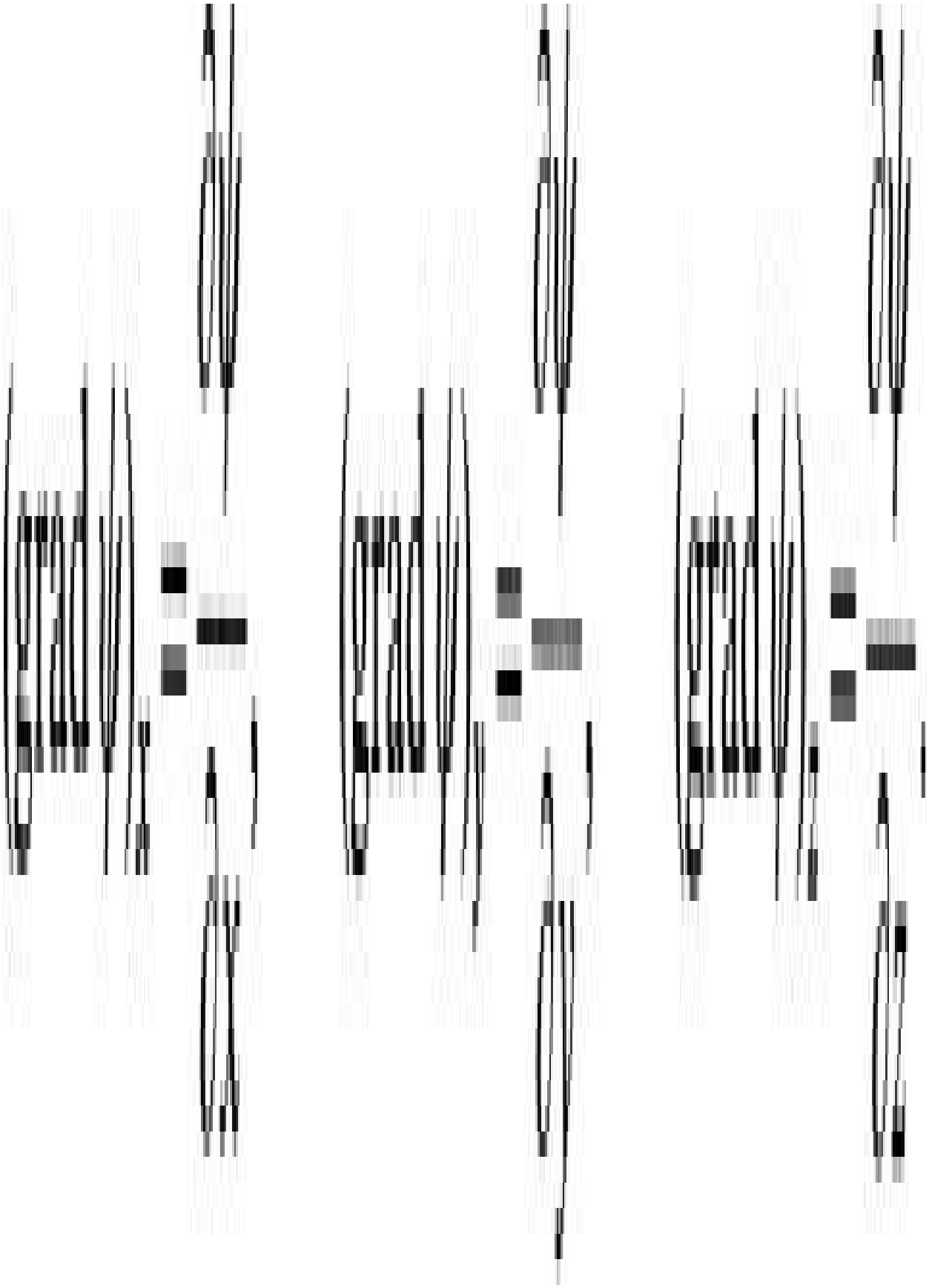
The components of curl F ≡ ∇ × F are

(B6)



For a scalar field $\psi(x, y, z)$ the components of grad $\psi = \nabla\psi$ are

(B7)



The Laplacian operator is

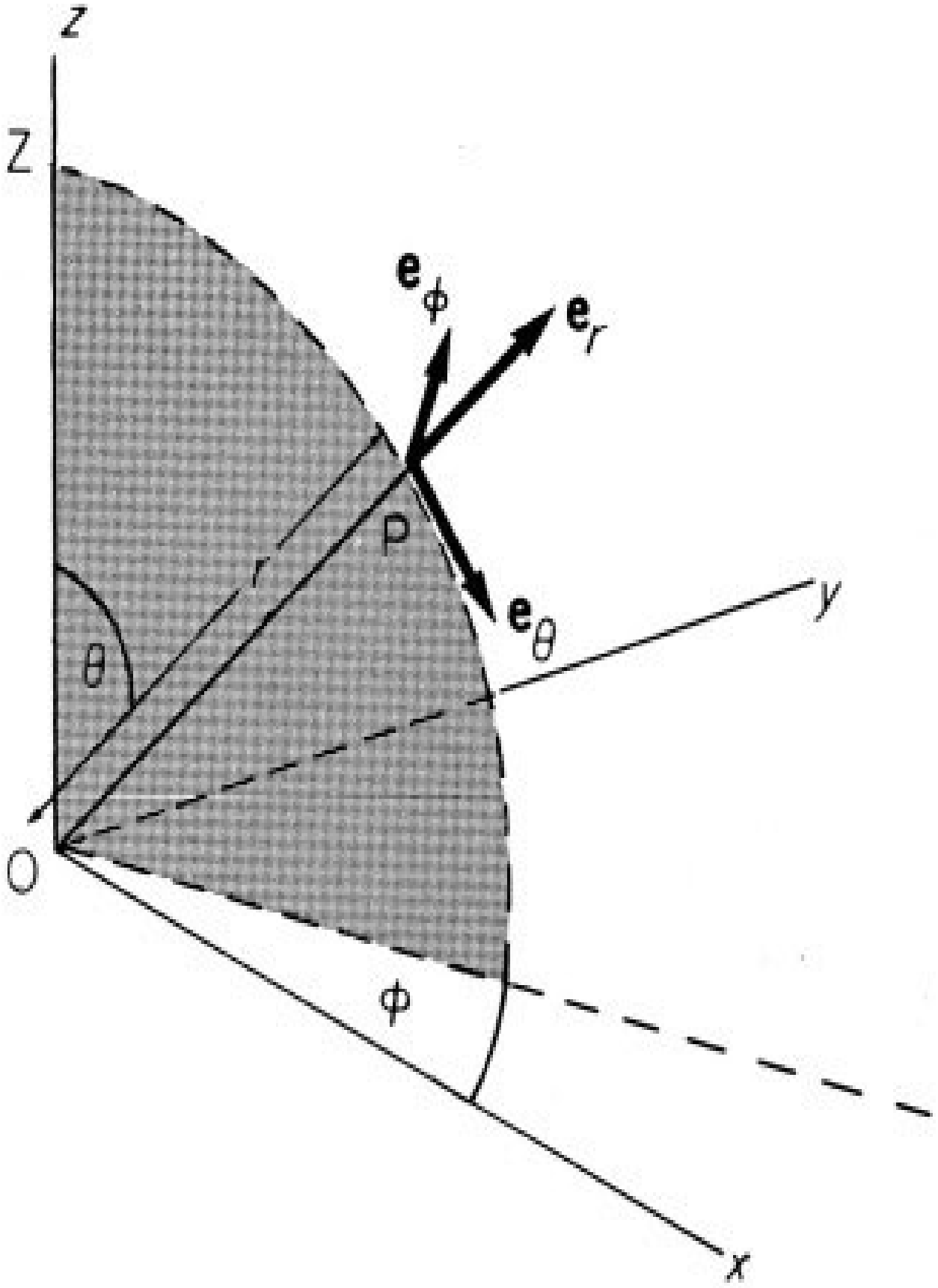
(B8)

Spherical polar coordinates

A point P has spherical polar coordinates (r, θ, ϕ), where r is the length OP, as indicated in Figure B.3; θ the angle between OP and the z-axis; and ϕ the angle between the x-axis and the plane OPZ containing P and the z-axis. In this coordinate system a vector field is resolved into components along directions which depend on where the field is being evaluated. At P a vector field F is resolved into components Fr, Fθ and Fϕ along the directions of the orthogonal unit vectors er, eθ and eϕ. The vector er points in the direction OP; eθ lies in the plane OPZ perpendicular to er and pointing in the direction of increasing θ; and eϕ is normal to the plane OPZ, pointing in the direction of increasing ϕ.

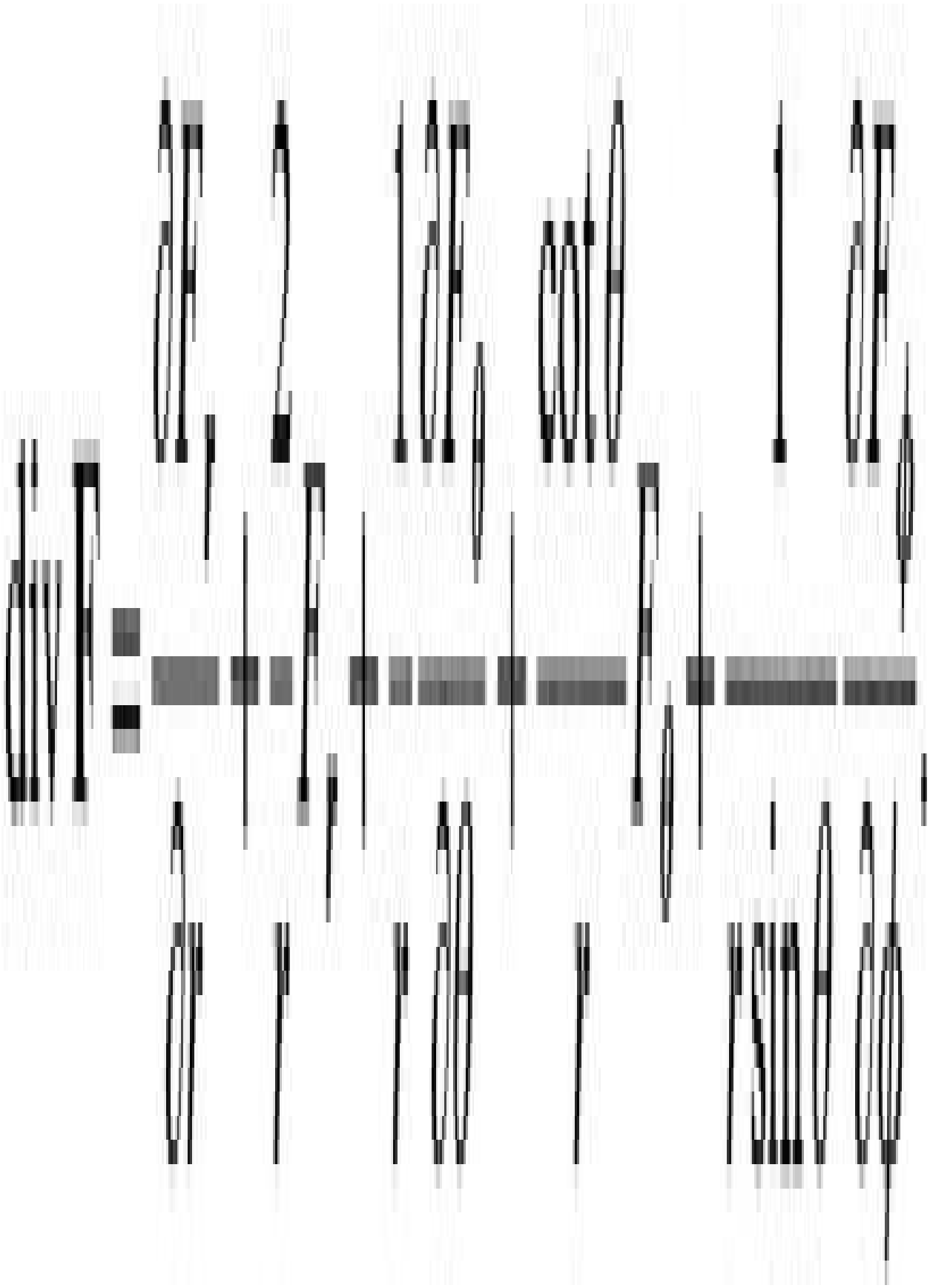
In polar coordinates the operator ∇ cannot be expressed in a simple way in terms of the unit vectors, and in working out the fields generated by the operator it is necessary to return to the definitions in Equations (B1), (B2) and (B3). The results are given below, without proof, for arbitrary fields F(r, θ, ϕ) and ψ(r, θ, ϕ).

Figure B.3.


z
Z
$\mathbf{e}_\phi$
$\mathbf{e}_r$
P
r
$\mathbf{e}_\theta$
y
θ
O
φ
x

The divergence of F is

(B9)



The components of curl F are

(B10)

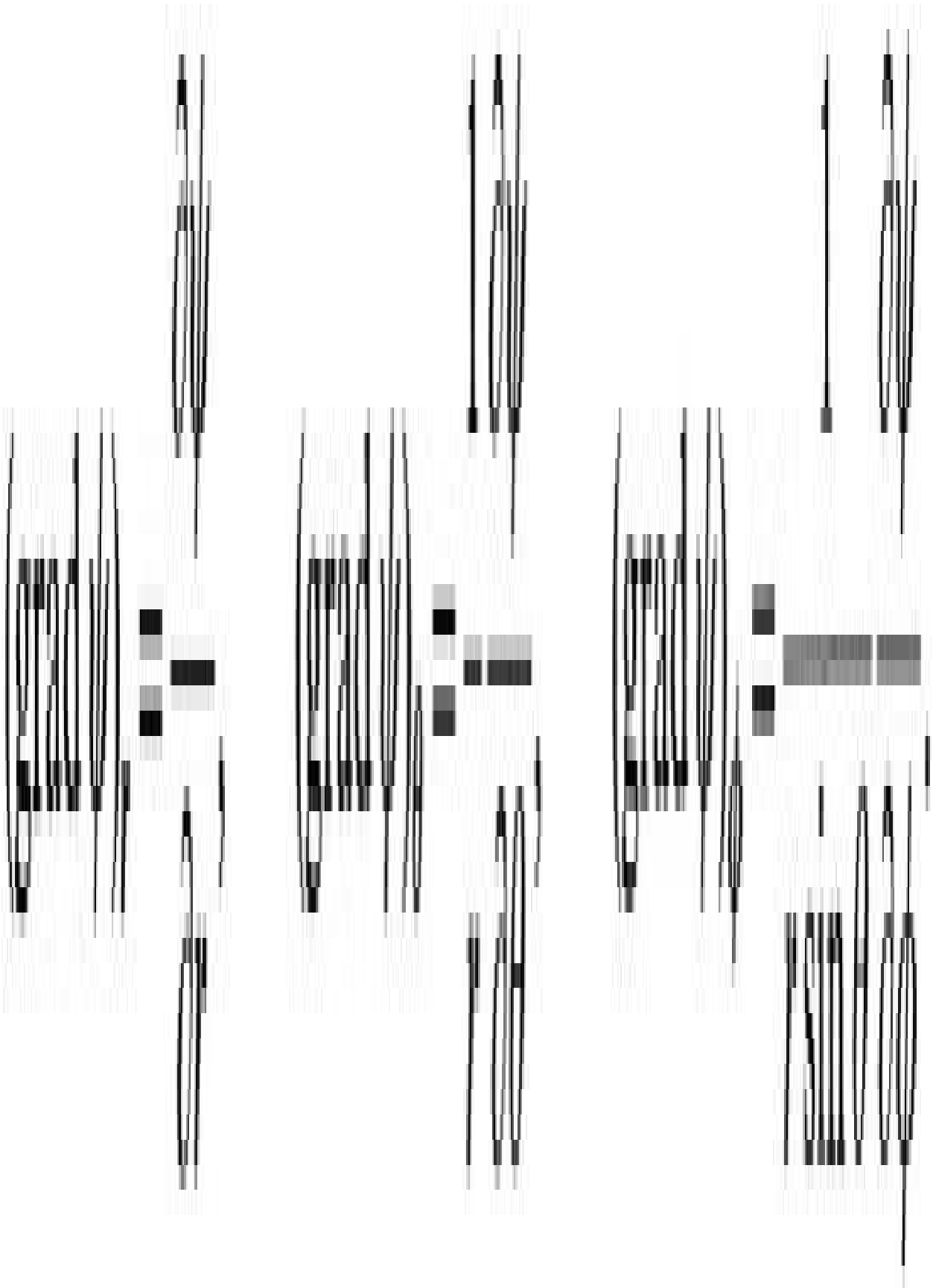
$$(\mathrm{curl}\,\mathbf{F})_r \equiv \frac{1}{r}\frac{\partial F_\phi}{\partial \theta} + \frac{\cot\theta}{r}F_\phi - \frac{1}{r\sin\theta}\frac{\partial F_\theta}{\partial \phi},$$

$$(\mathrm{curl}\,\mathbf{F})_\theta \equiv \frac{1}{r\sin\theta}\frac{\partial F_r}{\partial \phi} - \frac{\partial F_\phi}{\partial r} - \frac{1}{r}F_\phi,$$

$$(\mathrm{curl}\,\mathbf{F})_\phi \equiv \frac{F_\theta}{r} + \frac{\partial F_\theta}{\partial r} - \frac{1}{r}\frac{\partial F_r}{\partial \theta}.$$

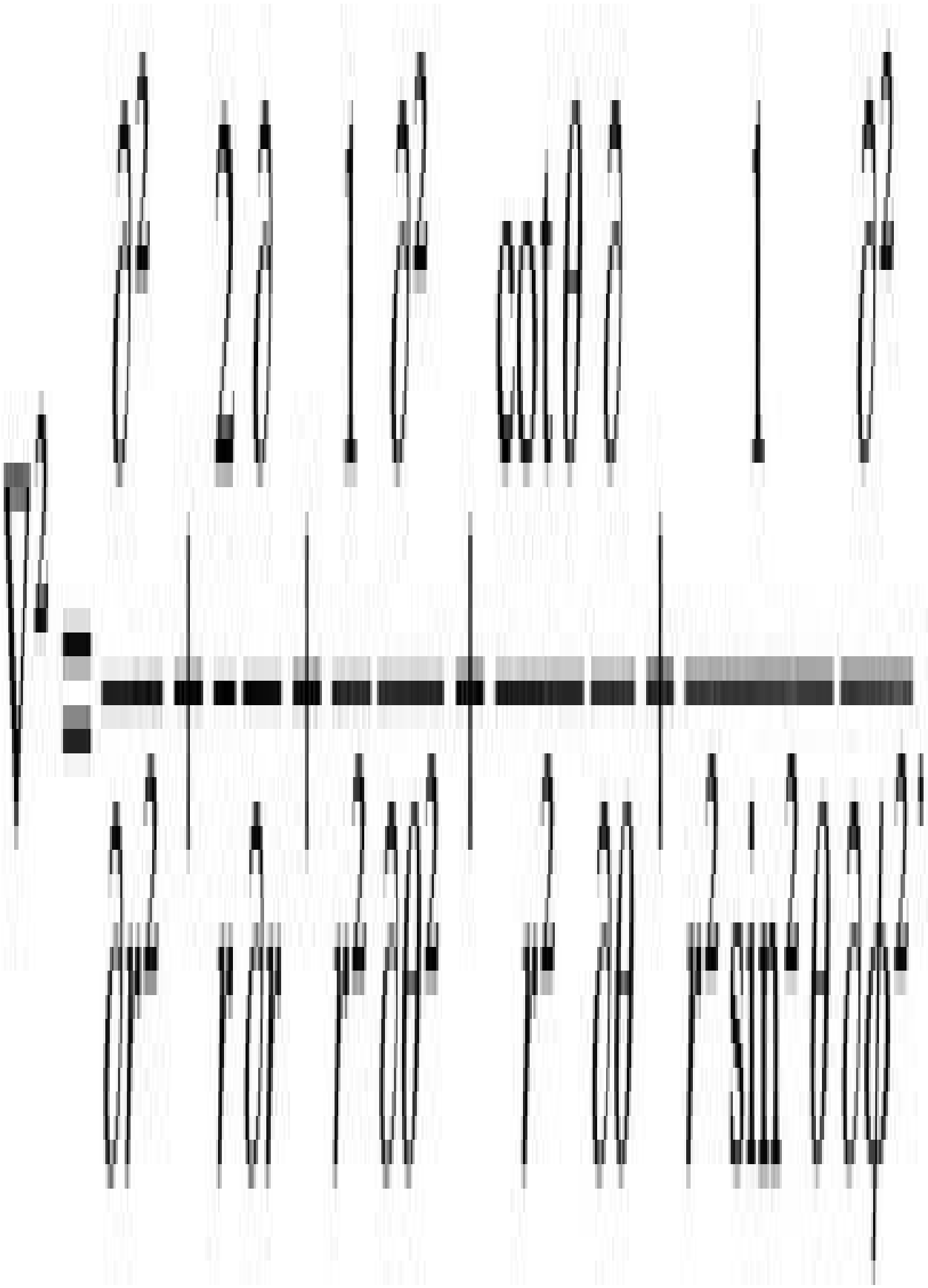
The components of grad ψ are

(B11)



The Laplacian operator is

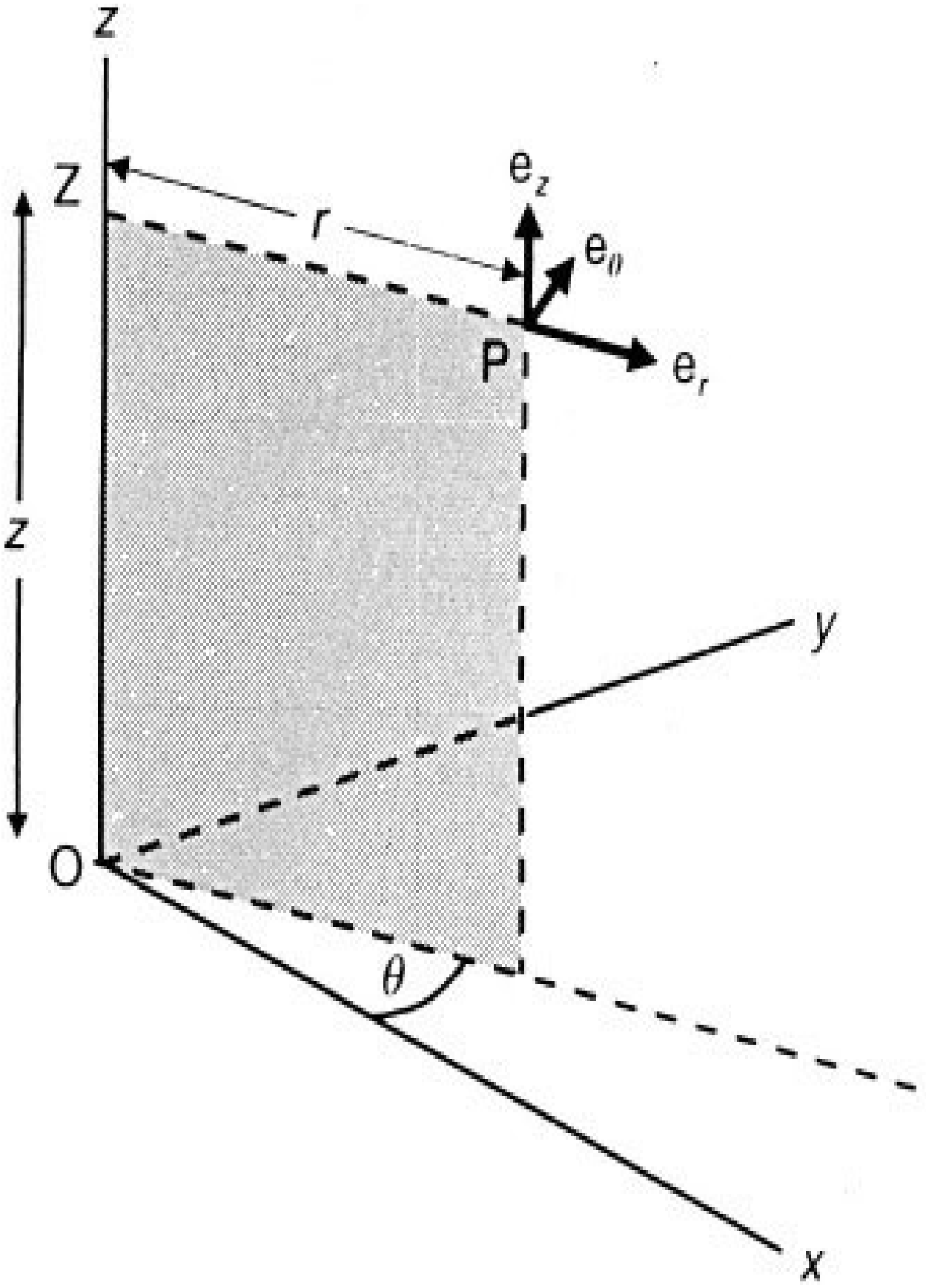
(B12)



Cylindrical polar coordinates

A point P has cylindrical polar coordinates (r, θ, z) as indicated in Figure B.4. Here r is the perpendicular distance from P to the z-axis, θ the angle between the

Figure B.4.


z
Z
r
e_z
$e_{//}$
P
e_r
z
y
O
θ
x

x-axis and the plane OPZ, and z the same as the Cartesian coordinate z. (Cylindrical polar coordinates are usually denoted (r, ϕ, z), but we have used θ for the azimuthal angle in order to avoid confusion with the electrostatic potential ϕ in the discussion of Laplace's equation on p. 87.) At P a vector field F is resolved into components Fr, Fθ and Fz along the directions of the orthogonal unit vectors er, ee and ez. The vector ez is parallel to the z-axis, er lies in the plane OPZ, perpendicular to ez, and eθ is perpendicular to the plane OPZ, pointing in the direction of increasing θ.

In this coordinate system, the divergence of F(r, θ, z) is

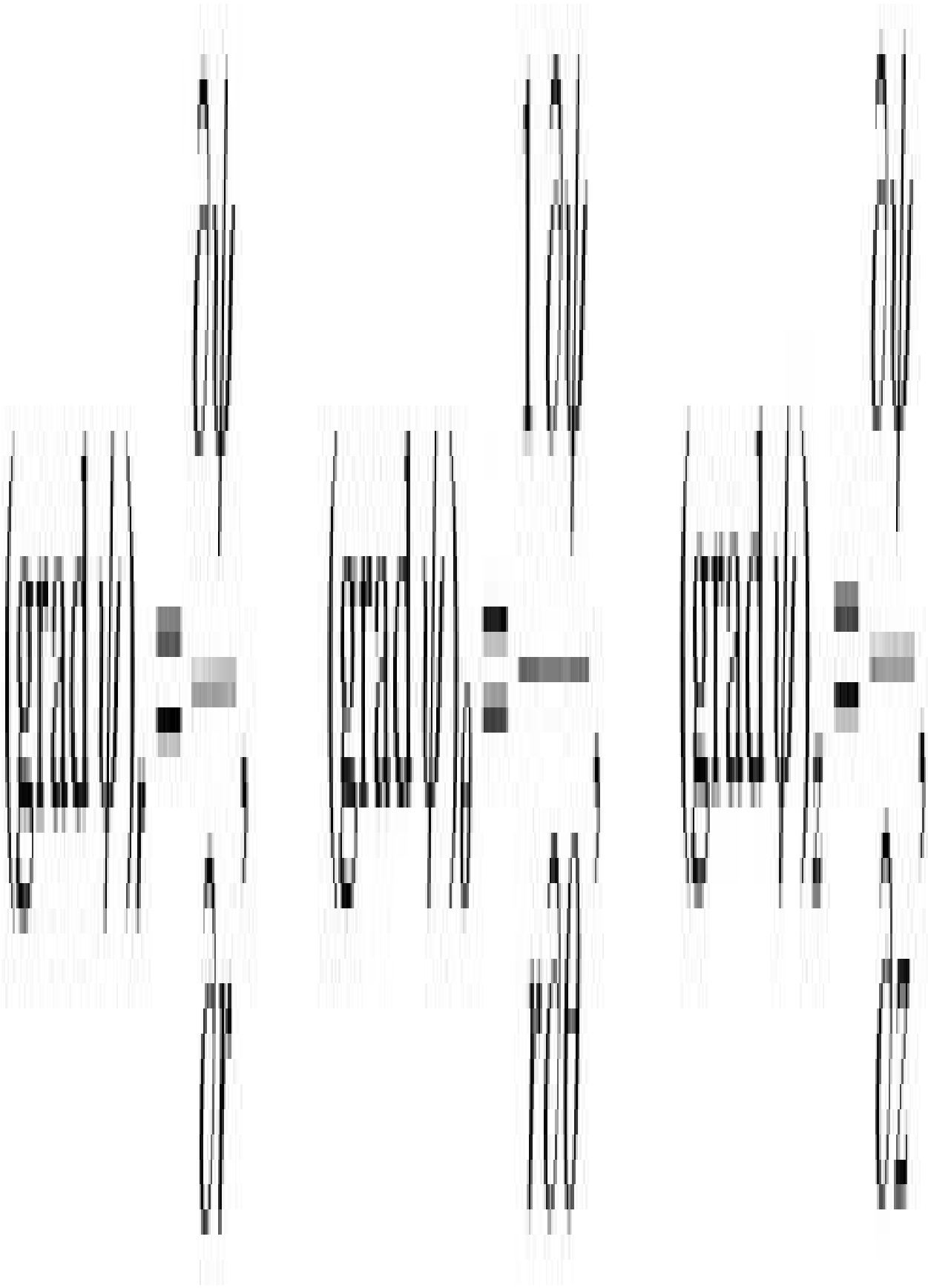
(B13)

The components of curl F are

(B14)

The components of the gradient of a scalar field ψ(r, θ, z) are

(B15)



The Laplacian operator is

(B16)

B.3 IDENTITIES

There are several equations involving fields and differential operators which hold for any smoothly-varying fields. These equations are mathematical identities which follow directly from the definitions of the differential operators. The identities required in the theory of electromagnetism are discussed in this section.

The divergence theorem

The total flux of a vector field F out of a closed surface S equals the volume integral of div F over the volume enclosed within S:

(B17)

The divergence theorem follows directly from the definition of div F given in section B.1, since $\int_s F \cdot dS$ equals the sum of the outward fluxes from all the infinitesimal volumes $d\tau$ making up the volume V.

Stokes' theorem

The line integral of a vector field F around a circuit s equals the flux of curl F through any surface S which spans s:

(B18)

The line integral is taken in the sense related by the right-hand rule to the direction of the field. Stokes' theorem follows directly from the definition of curl given in section B.1, since $\oint_s \mathbf{F} \cdot d\mathbf{l}$ is equal to the sum of the line integrals around the perimeters of the mosaic of elementary surfaces dS which make up S.

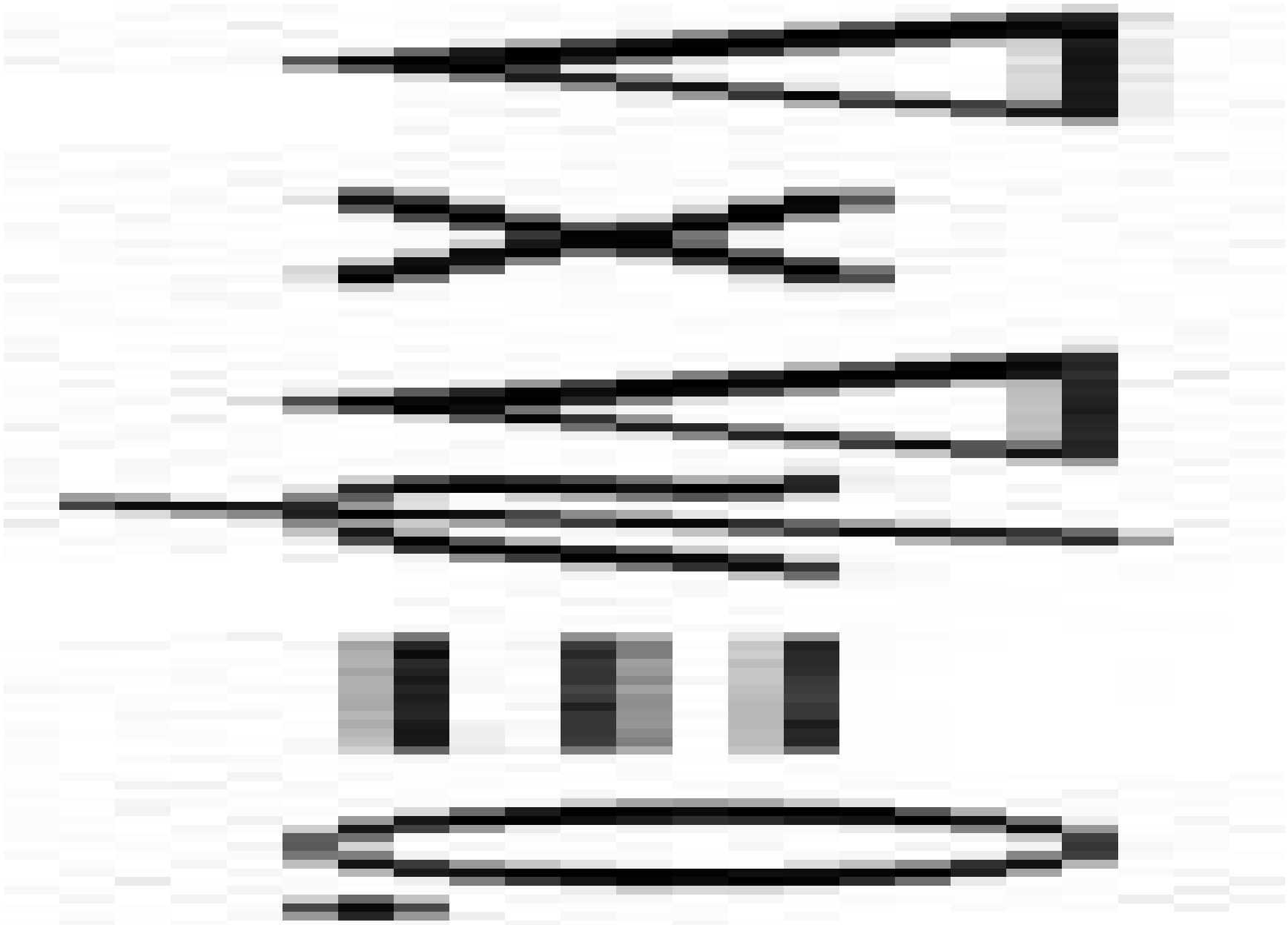
Identities involving products of the vector operator ∇

The operator ∇ is itself a vector, and obeys the same rules for forming vector products as any other vector. You must, however, be careful when inserting V into vector relations, because ∇ is a differential operator as well as a vector, and it must be allowed to operate on all fields to the right of it. The identities given below follow from the rules relating to vector products:

(i) For any scalar field ψ,

(B19)

or



(This identity is true because the operator $\nabla \times \nabla$ is zero, since it is the vector product of a vector with itself.)

(ii) For any vector field F, div curl F ≡ 0, or

(B20)

(Scalar triple products are zero when any two of the vectors in the product are the same. But note! As was mentioned in section 4.5.1, F·curl F ≡ F · ($\nabla \times$ F) is not necessarily zero. This is because the differential operation is carried out on only one of the two fields F, which are therefore not equivalent.)

(iii) curl curl F ≡ grad divF − ∇^2F, or

(B21)

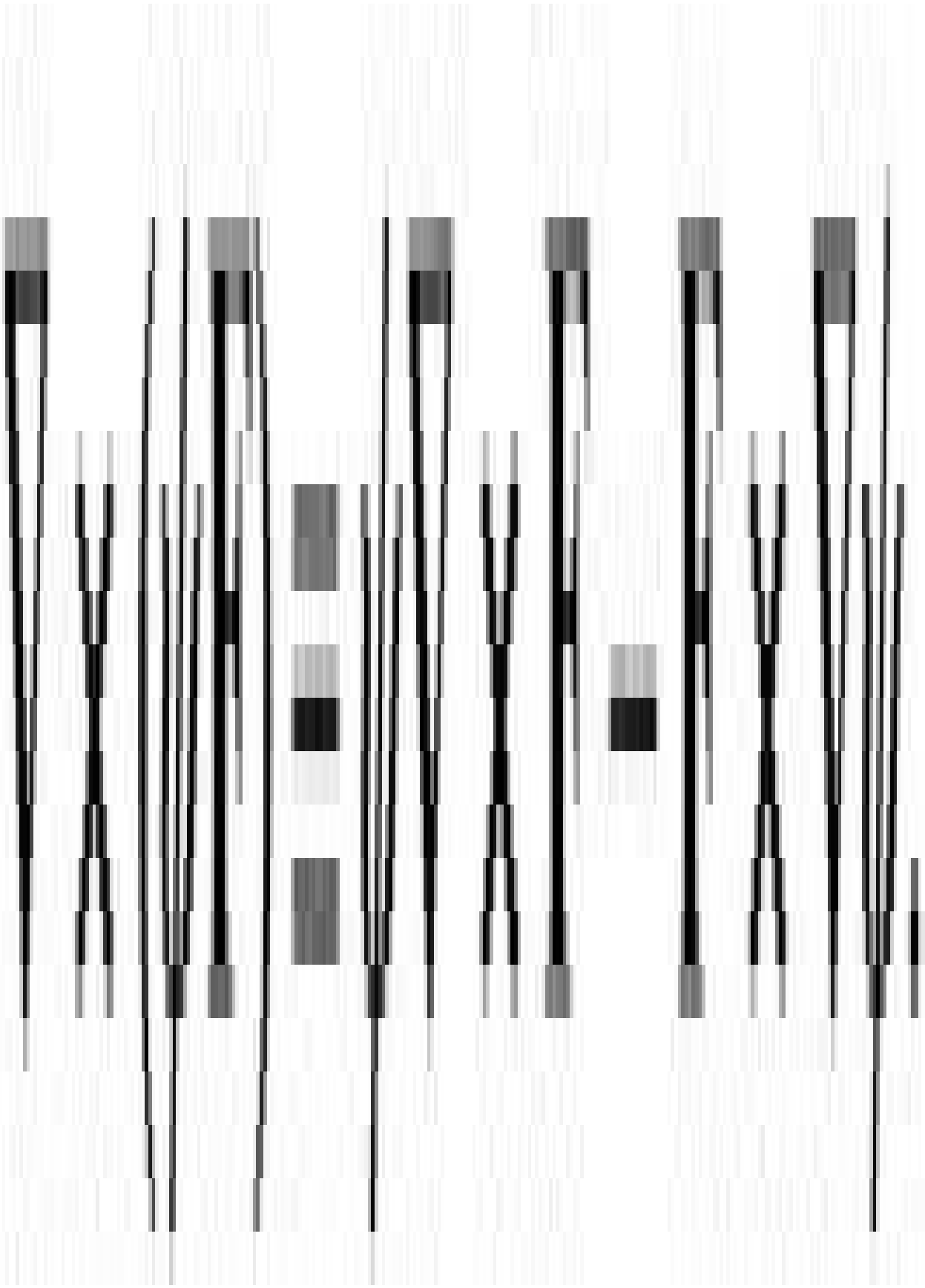
Identities including products of two fields

Here again the identities follow from the application of the rules of vector algebra. One must be careful about the order in which the fields and operators are written, specifying clearly which fields are operated on by ∇

(i) $\operatorname{div}(\psi F) \equiv \psi \operatorname{div} F + F \cdot \operatorname{grad} \psi$, or

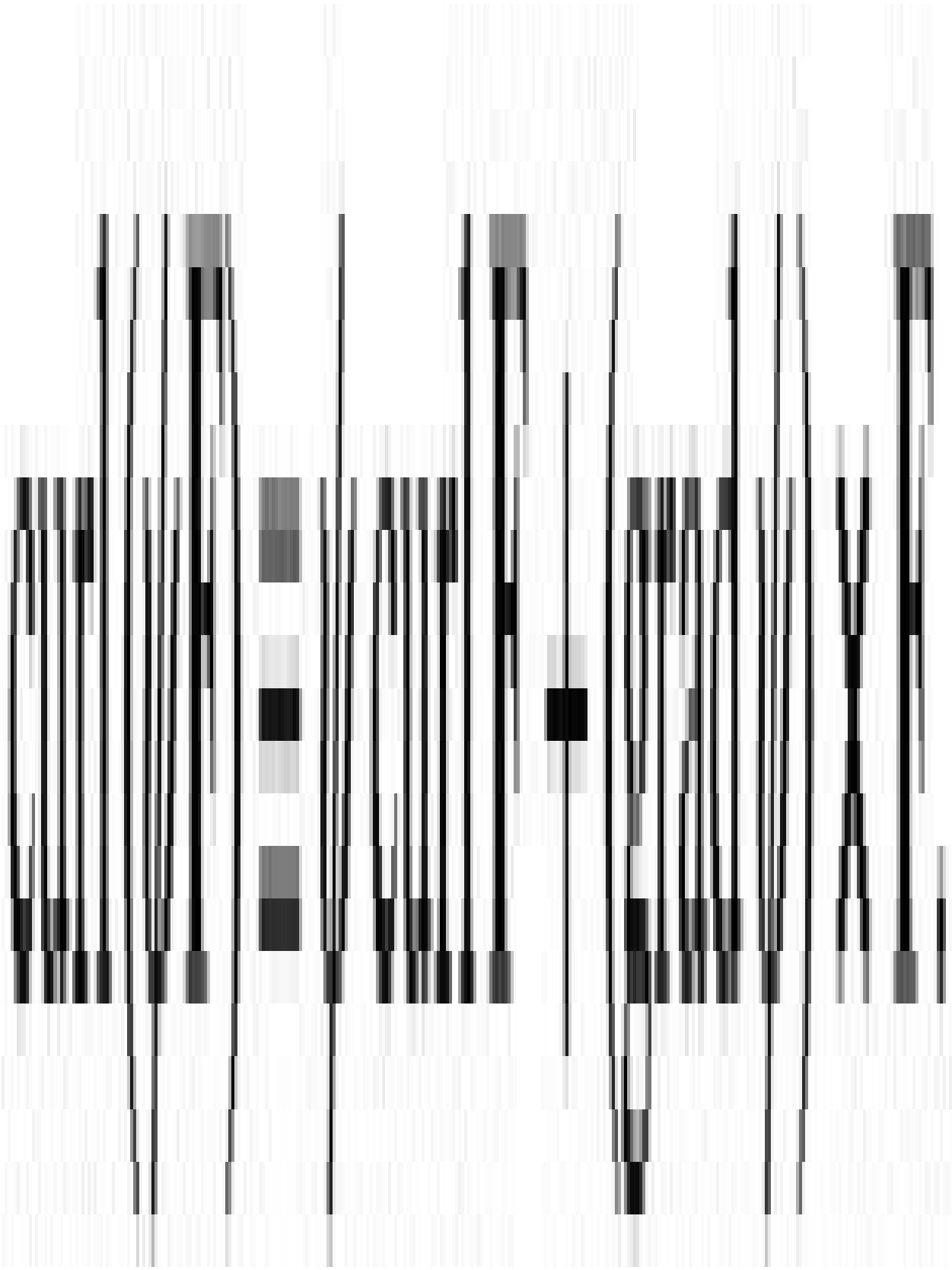
(B22)

(ii) $\operatorname{curl}(\psi F) \equiv \psi \operatorname{curl} F - F \times \operatorname{grad} \psi$, or



This identity could also be written as

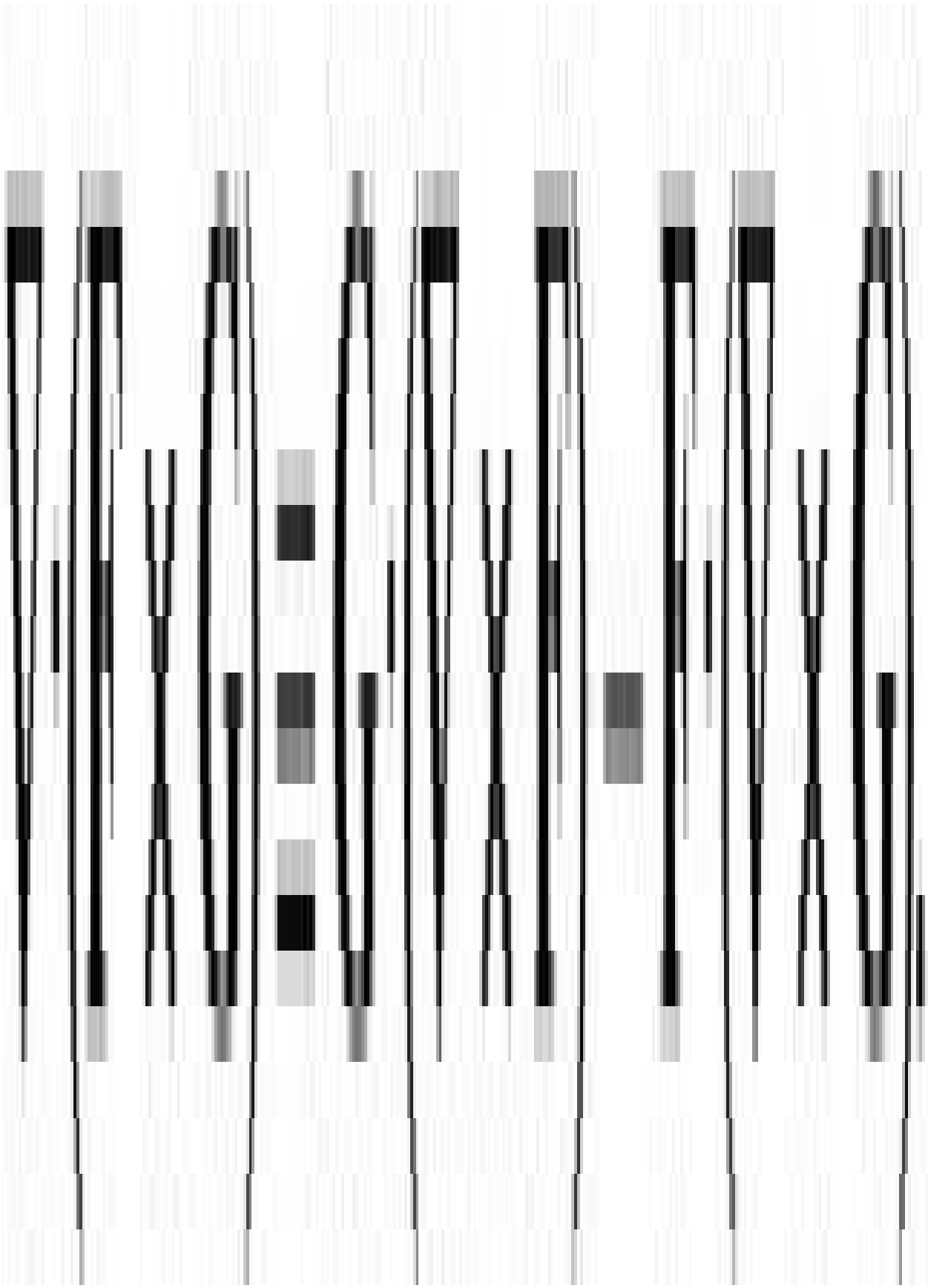
(B23)



We have preferred to reverse the order of the second term on the right-hand side, with a consequent change of sign. In this notation, the operator is then allowed to operate on everything to its right in each term.

(iii) div $(F \times G) \equiv G\text{-curl } F - F\text{-curl } G$, or

(B24)



The scalar triple product $a \cdot (b \times c)$ represents the volume of a parallelopiped with sides a, b and c Its value is independent of the way it is written except that there is a change of sign if the cyclic order of a, b and c is altered. The identity written above is just a rearrangement of the components of the triple scalar product; there are two terms on the right-hand side because ∇ must be allowed to operate on both the fields F and G.

APPENDIX C

The derivation of the Biot-Savart law

In this appendix we derive the Biot-Savart law, starting from the differential form of Ampère's law, Equation (4.47).

From Equations (4.47) and (4.49), we obtain

(C1)

The quantity curl curl A may be rewritten using the identity (B.21);

(C2)

where ∇^2 is the Laplacian operator; $\nabla^2 A$ is the vector whose three components in Cartesian coordinates are

$$\frac{\partial^2 A_x}{\partial x^2} + \frac{\partial^2 A_x}{\partial y^2} + \frac{\partial^2 A_x}{\partial z^2}$$

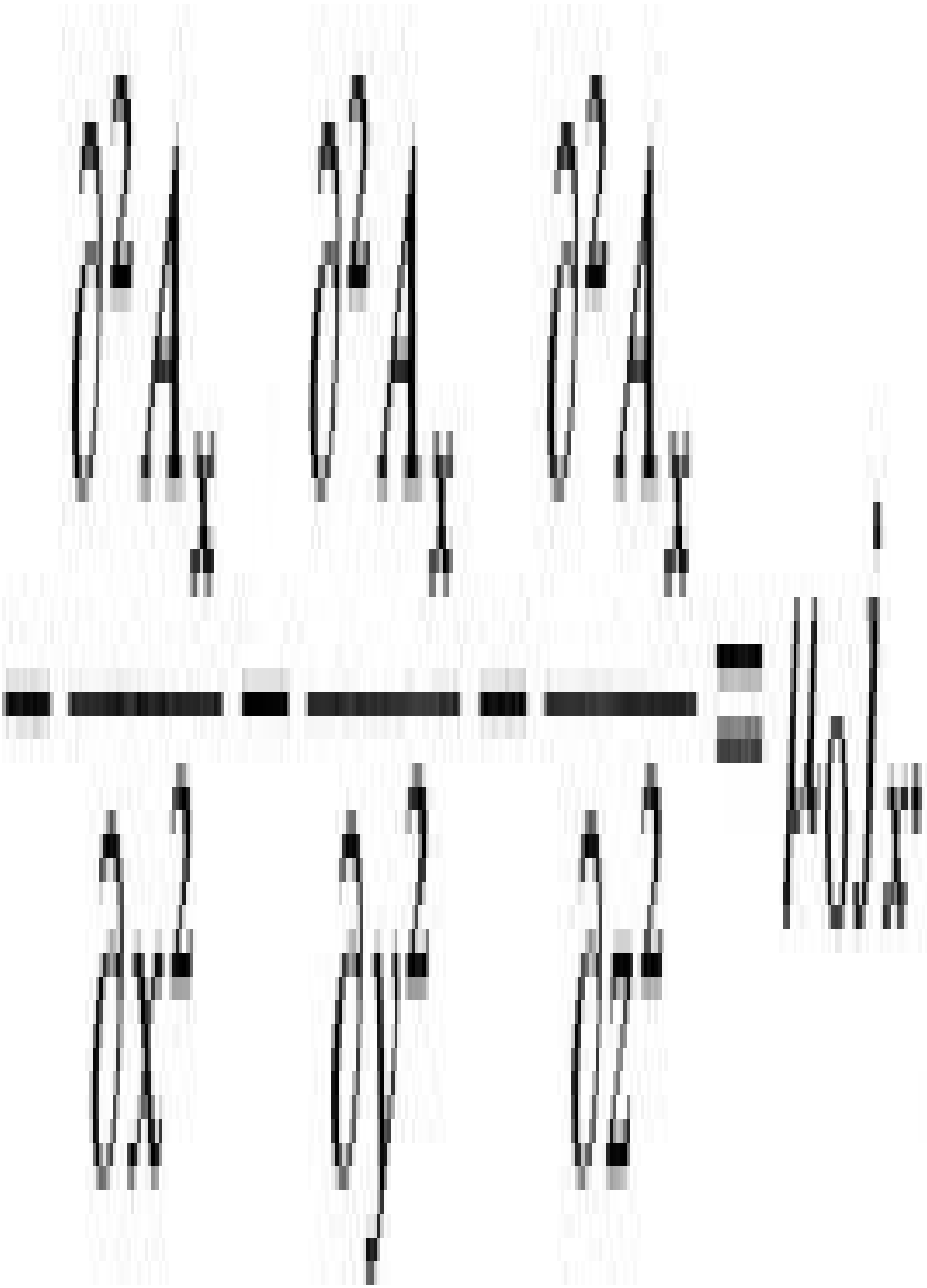
$$\frac{\partial^2 A_y}{\partial x^2} + \frac{\partial^2 A_y}{\partial y^2} + \frac{\partial^2 A_y}{\partial z^2}$$

$$\frac{\partial^2 A_z}{\partial x^2} + \frac{\partial^2 A_z}{\partial y^2} + \frac{\partial^2 A_z}{\partial z^2}.$$

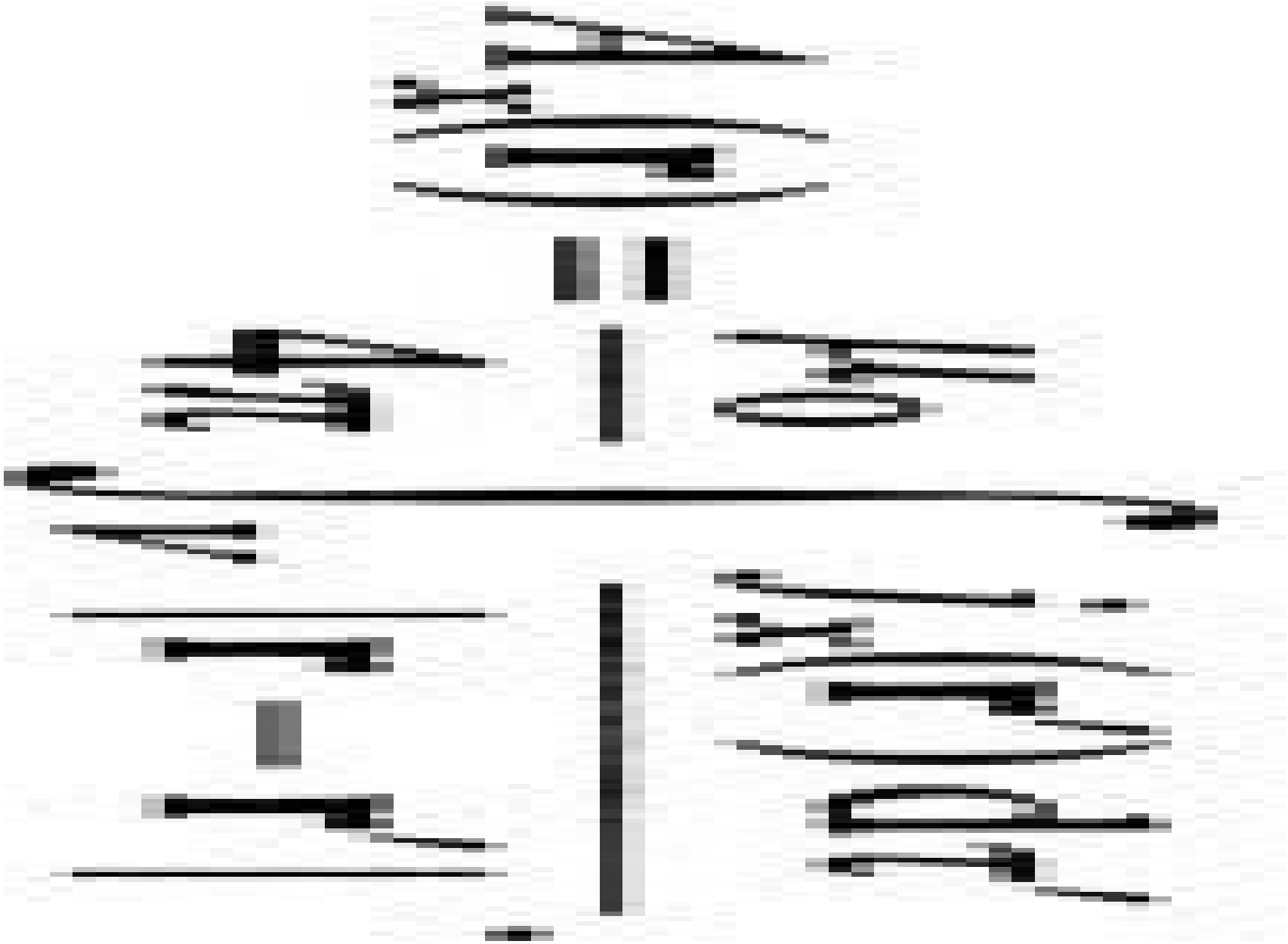
With the choice div A = 0, Equations (C1) and (C2) reduce to the equation

(C3)

This equation has three component equations. For example in Cartesian coordinates the x-component is



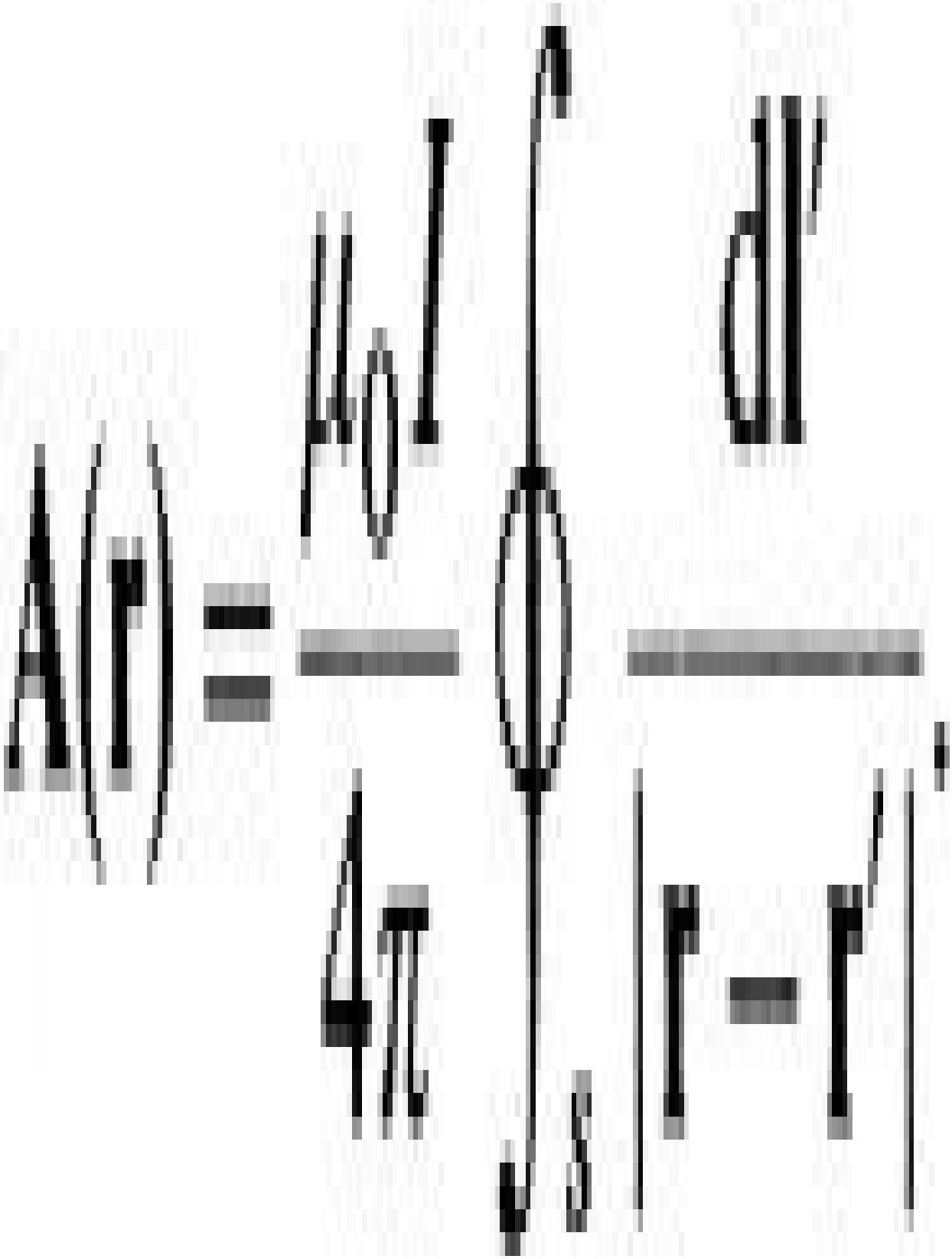
Each one of the Cartesian components is an equation similar to Poisson's Equation (3.2), and the solution to the equation for Ax for example is similarly



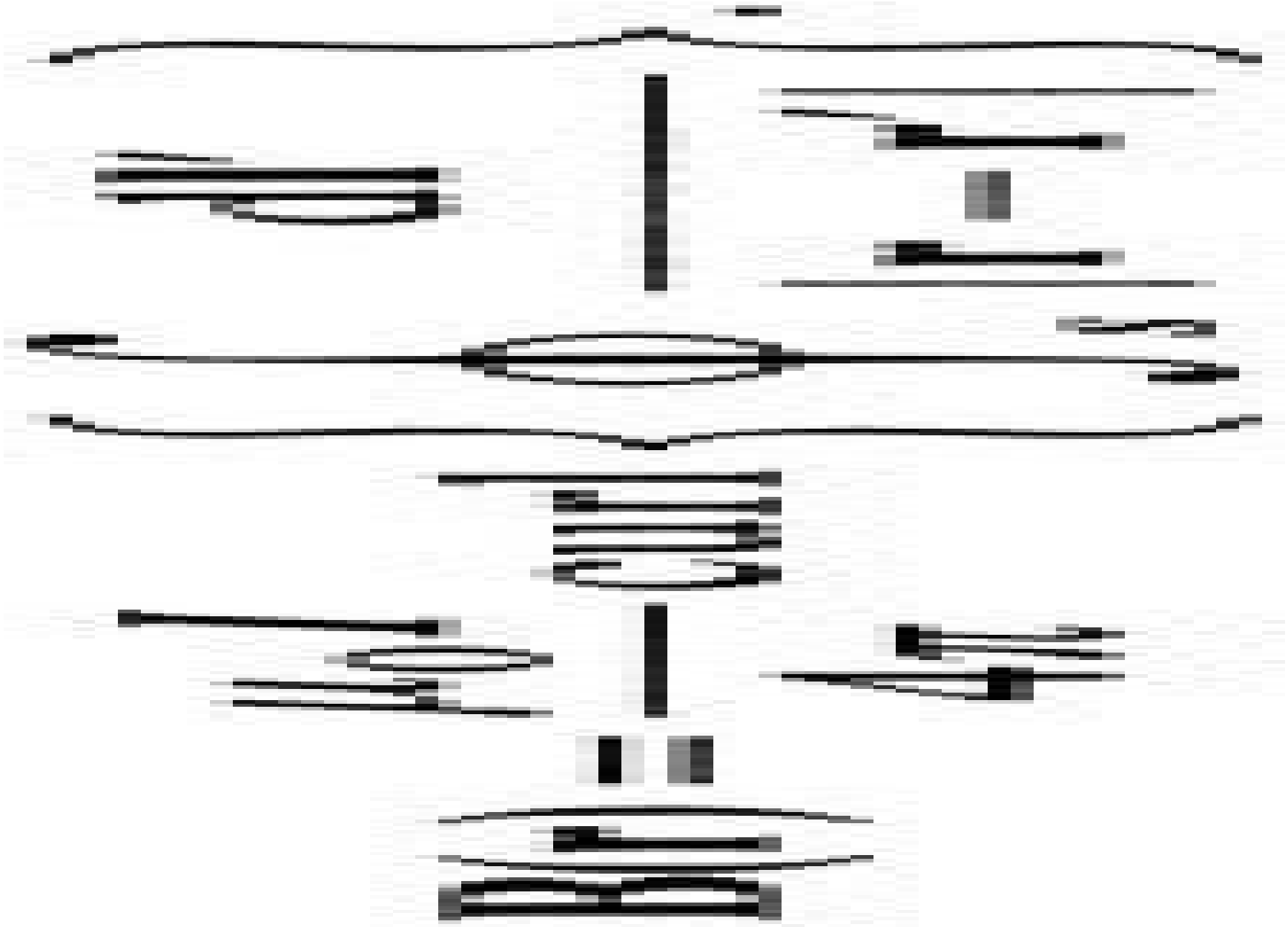
The general solution to Equation (C3) is then

(C4)

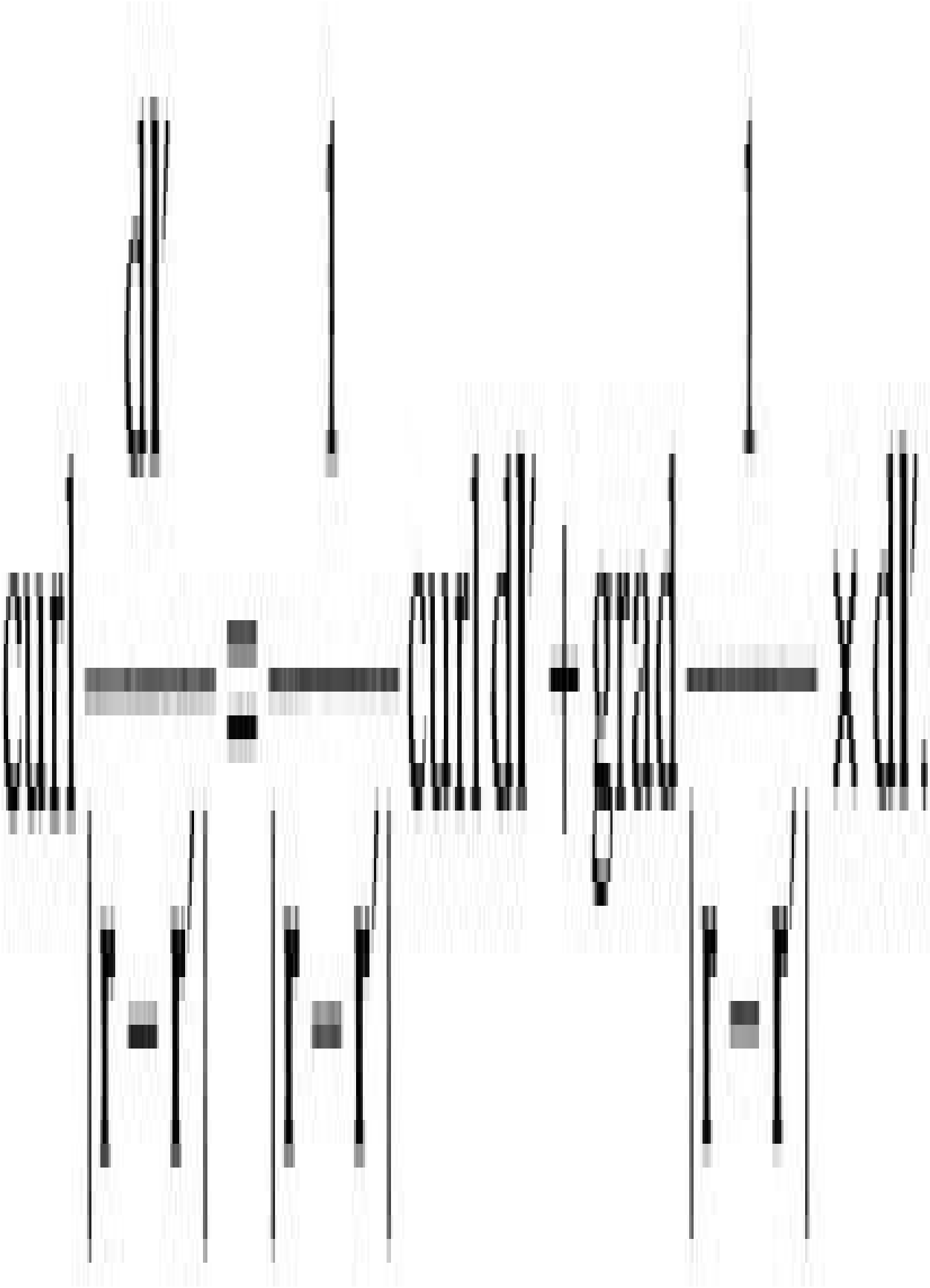
When the current is flowing in a thin conductor the current density j(r) is zero everywhere except along the conductor. The term j(r’) dτ’ in the integrand above can then be replaced by the term I dl′, where I is the current and the vector dl′ is in the direction of the current in the conductor at the point with position vector r′. The integral over the volume V containing the current distributions now becomes an integral over the closed length s of the conductor, and we have



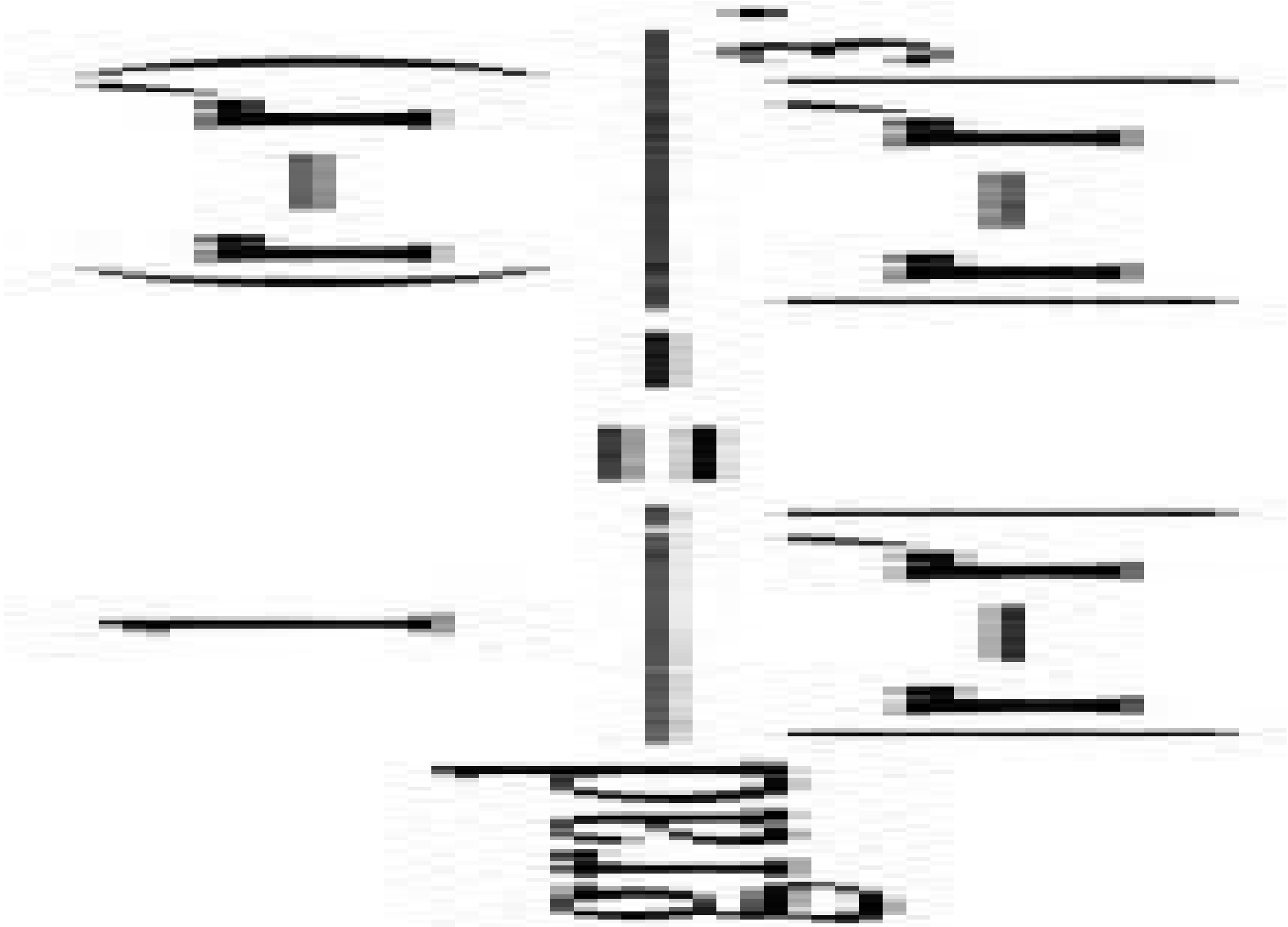
Hence



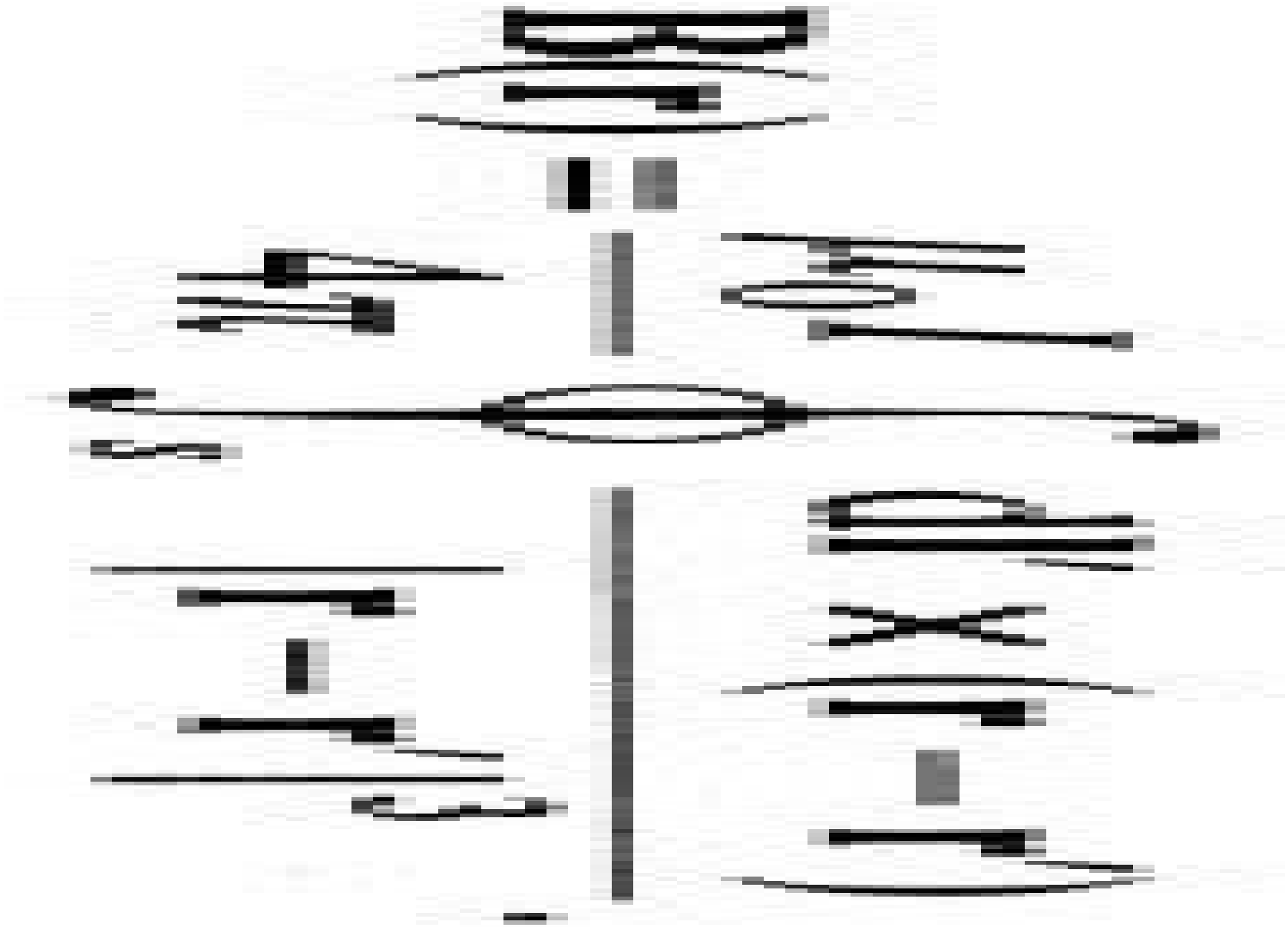
The curl operation is with respect to the coordinates of the point with position vector r, hence the operator can be taken inside the integral. Applying identity (B.23) we find that



But curl dl’ = 0, since the differentiations are with respect to r, and also



Hence finally



This is the Biot-Savart law used in section 4.4.2.

Solutions to problems

PROBLEMS 1

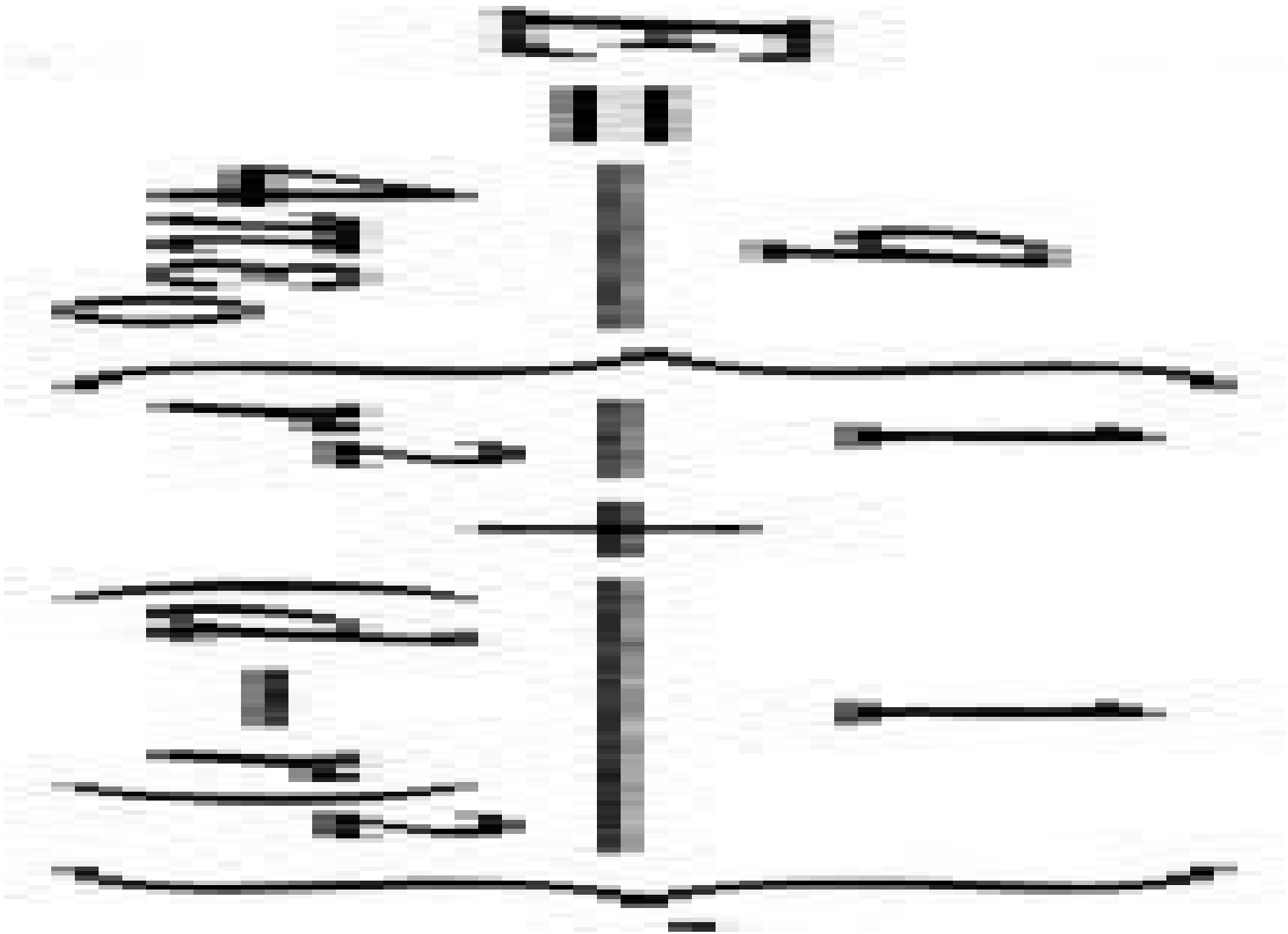
1.1 The force on each charge is $1.6 \times 10{-}2$ newtons. The field is zero at the centre of the triangle, and has a magnitude 12 000 V/m at the centre of each side.

1.3 10^5 V.

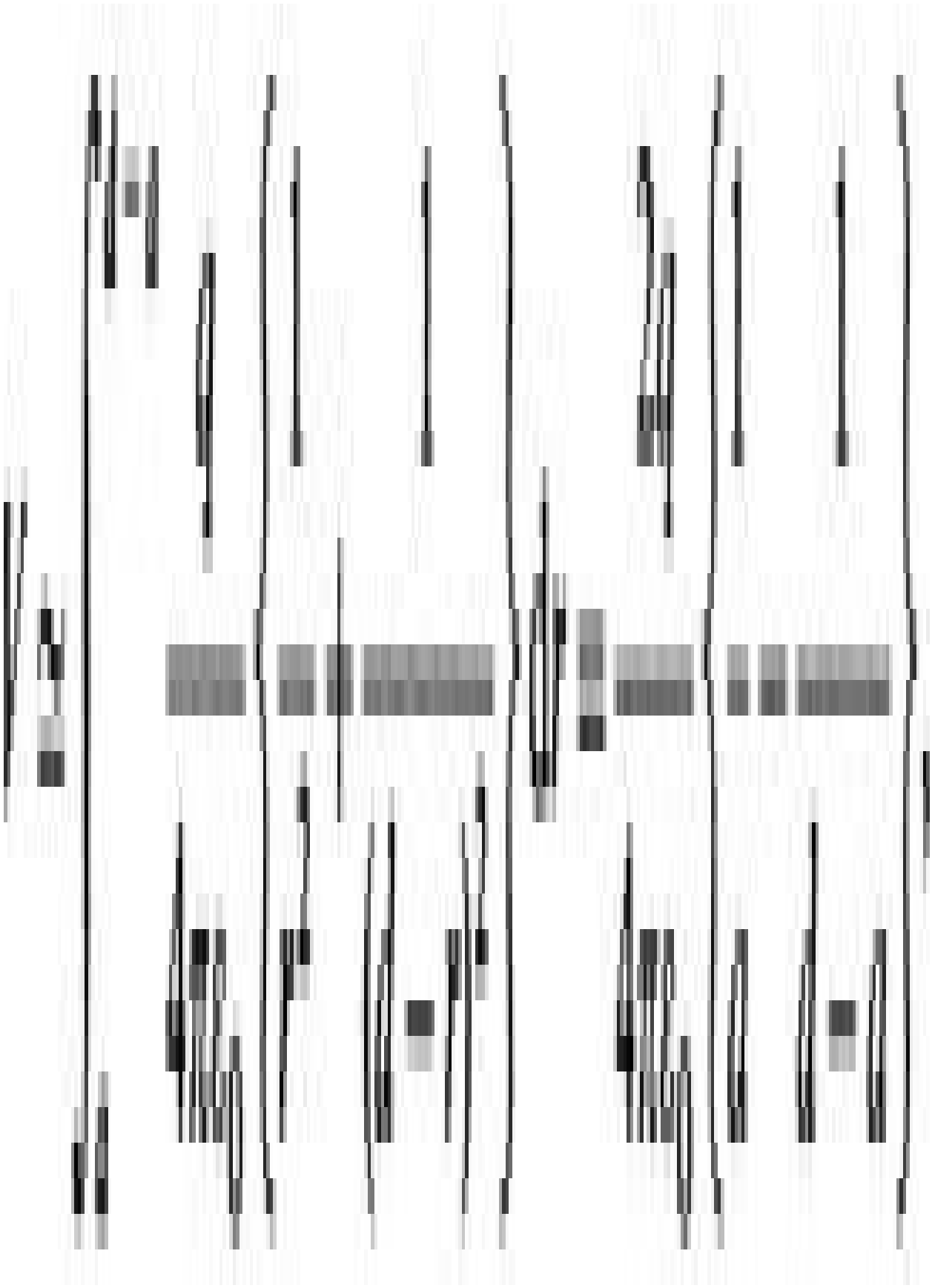
1.4 31.1 MV (MV = megavolt = 10^6 volts). When the terminal is at this voltage the intershield must be at 18.8 MV to prevent breakdown.

1.5 The capacitance is $4\pi\varepsilon 0/(1/a - 1/b)$. When the radius b of the outer sphere becomes large, the capacitance tends to the limiting value $4\pi\varepsilon 0a$. The same limit is reached when the sphere is surrounded by a distant conductor of any shape: this limit is often referred to as the capacitance to earth of an isolated sphere.

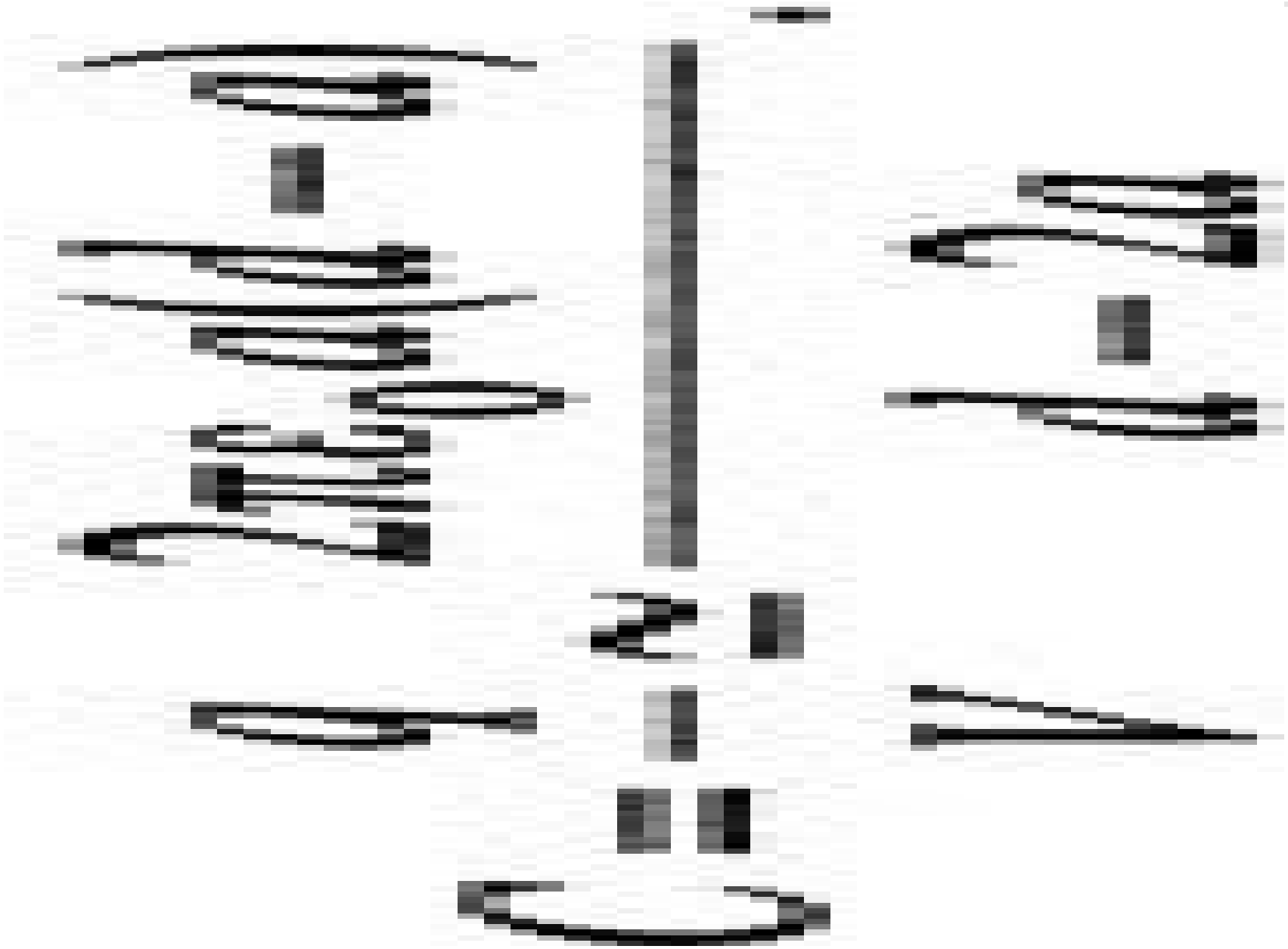
1.6 On the line joining two charges ± q separated by distance d the magnitude of the field a distance r from one of them is



This field points towards the negative charge. The potential difference V between spheres of radius a can be found from the field on the line joining their centres; if the field is almost the same as for the point charges, then



and the capacitance



1.7 The electrostatic potential energy of a uniformly charged sphere of radius R carrying a total charge Q is $Q^2/(4\pi\varepsilon_0 R)$. For a uranium nucleus Q = 92e and R = 10−14 m, and this energy amounts to 750 MeV. If the nucleus undergoes fission and splits into two equal spherical fragments, then the electrostatic potential energy of the pair of fragments is about 470 MeV. The reduction of 280 MeV is more than the energy actually released in uranium fission (just over 200 MeV) because the increased surface area of the two fragments is energetically unfavourable.

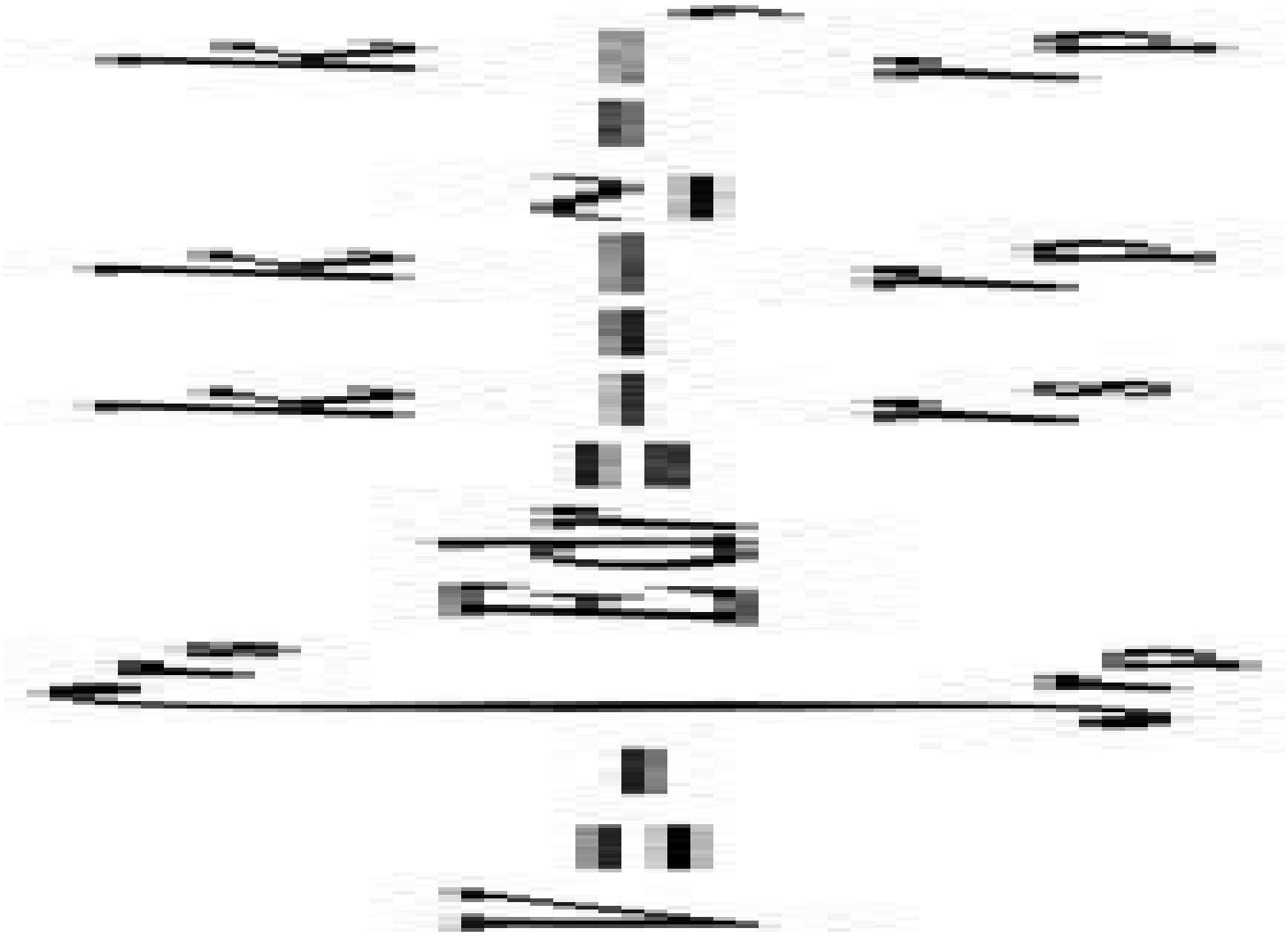
1.8

1.9 The quadrupole potential for r a is $(qa^2/4\pi\varepsilon_0 r^3)(3\cos^2\theta - 1)$. The dipole potential and the quadrupole potential are two of a series of multipole potentials, which are successively proportional to higher powers of (1/r). The multipole potentials may be used as the basis of an expansion for any potential, in much the same way as powers of sines and cosines form the basis of a Fourier expansion.

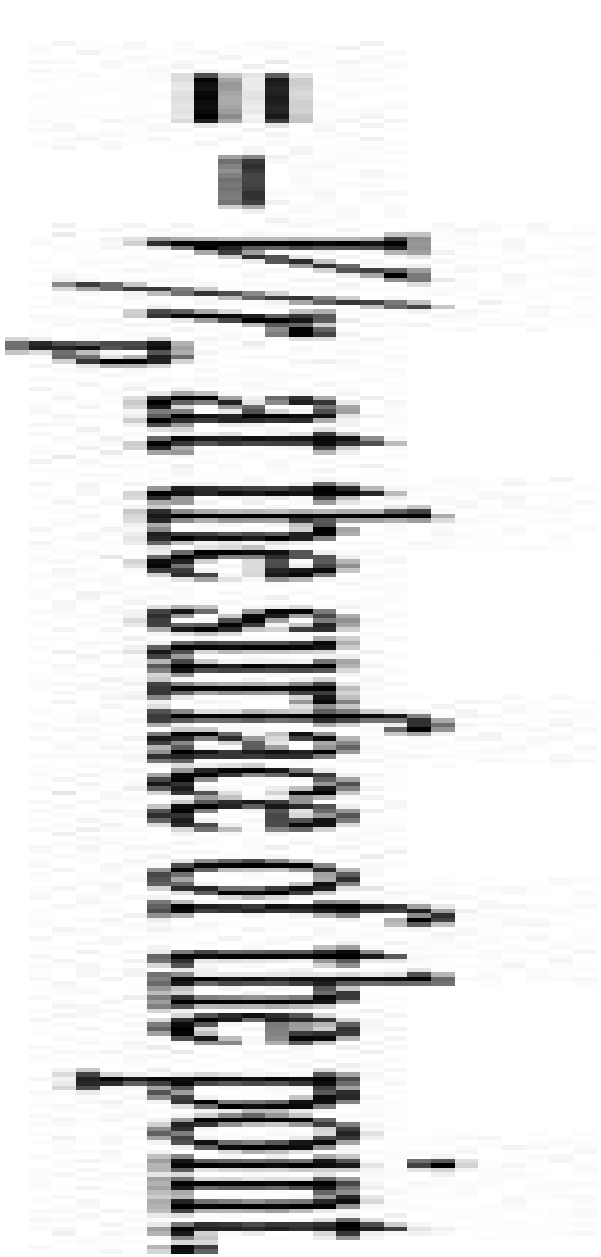
1.10 The force per unit area between the central electrodes is $F = \varepsilon_0 E^2 = \varepsilon_0 (V/d)^2 = \rho g h$. Hence . Note that this result is independent of edge effects and that no electrical measurements are required. This kind of voltmeter is used for absolute voltage calibration, and is capable of a precision of better than 1 ppm.

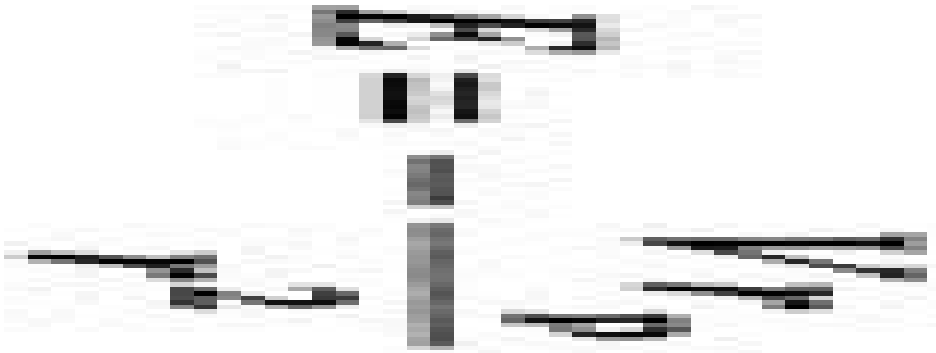
1.12 The field lines are straight, going radially from the point to the screen. An object on the point thus appears on the screen magnified by the ratio of the radii of curvature, namely 10^7.

If there is no charge between point and screen, the electric field at a distance r from the centre is proportional to $1/r^2$, i.e. $E = k/r^2$. The potential difference between the point (at rp) and the screen (at rs) is



and





Since rp is 10−8 m, E reaches the value of 10^8 V/m needed to cause field emission when the screen potential is only one volt. In practice much higher potential differences are required to produce a large enough current of energetic electrons to give a visible image on the screen—typically a potential difference of about 10 kV would be applied.

Although the simple microscope described here does have a very high magnification, it is of limited application. However, field emission electron sources are used in sophisticated electron microscopes, which include lenses to focus electron beams. The electron beam from a field emission source has a very high brightness, which is particularly useful in scanning transmission electron microscopy.

1.13 For the beam perpendicular to the grids, the electric field is parallel to the beam, and must stop the beam altogether if the beam is not to emerge. To slow the negatively–charged beam, the field should point in the direction of motion, so that a potential of 0 V will just stop the beam. When the beam is incident at 45°, only the component of velocity parallel to the field is changed, and a grid at 1 kV will stop the beam from emerging.

PROBLEMS 2

2.1 The final potential difference is $2V/(\varepsilon + 1)$.

2.2 (i) charge $Q = CV$, energy stored $U = CV^2$, work done by battery $W = CV^2$

(ii) $Q = \varepsilon CV$, $U = \varepsilon CV^2$

$\Delta Q = (\varepsilon - 1)CV$, $\Delta U = (\varepsilon - 1)CV^2$, $\Delta W = V\Delta Q = (\varepsilon - 1)CV^2$

The battery does work in pulling the dielectric into the capacitor as well as in storing electrostatic energy.

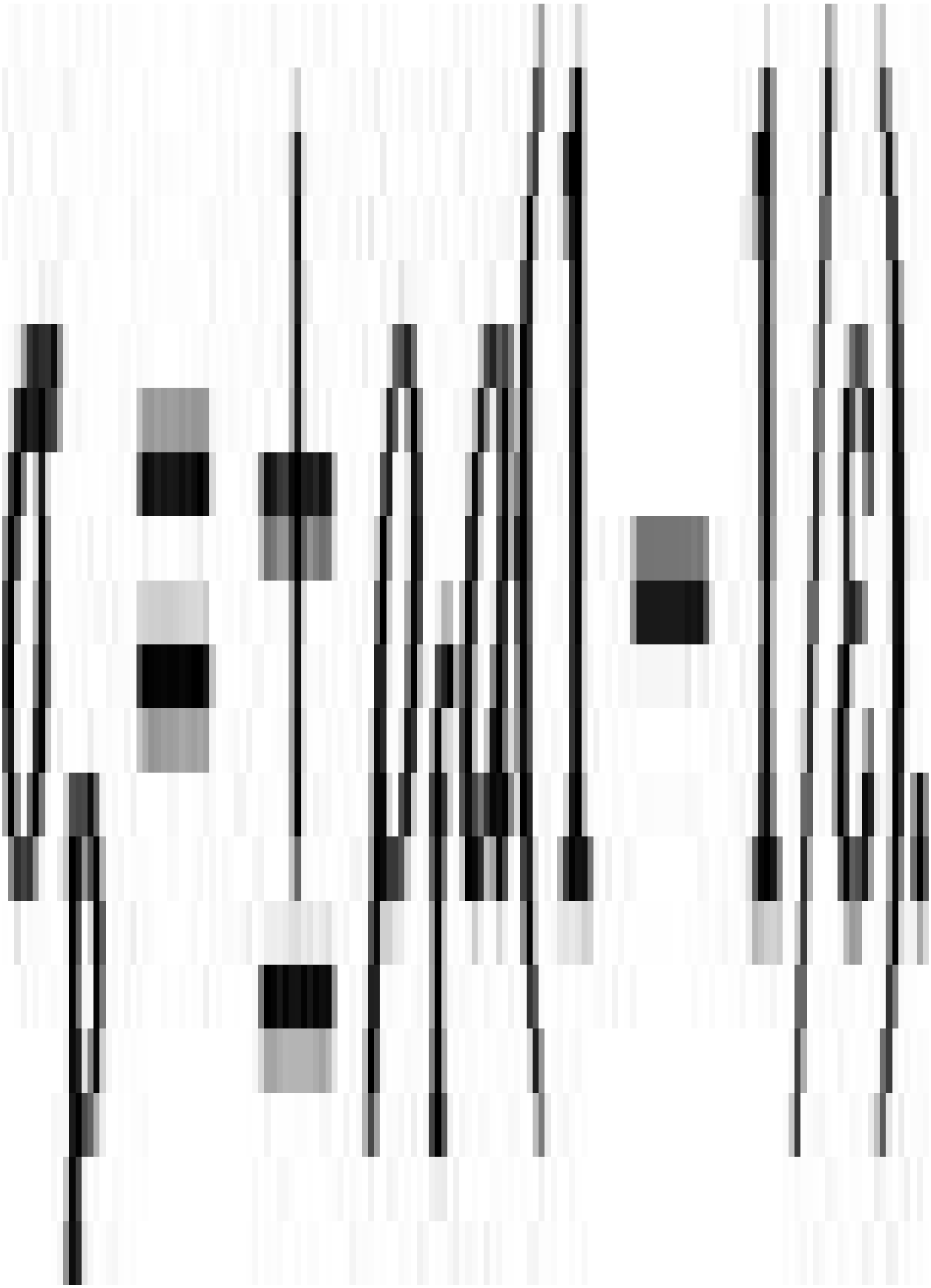
(iii) $Q = \varepsilon CV$, $U = Q^2/C = \varepsilon^2 CV^2$ i.e. more than the total amount of work [$CV^2 + (\varepsilon - 1)CV^2$] done by the battery during the sequence of operations. The difference $(\varepsilon - 1)^2 CV^2$ represents work done by gravity.

2.3 If the capacitor is entirely filled with polycarbonate, the thickness of the sheet is about 5×10^{-6} m. When 50 V is applied across the sheet, it experiences a compressional pressure of 800 N/m^2.

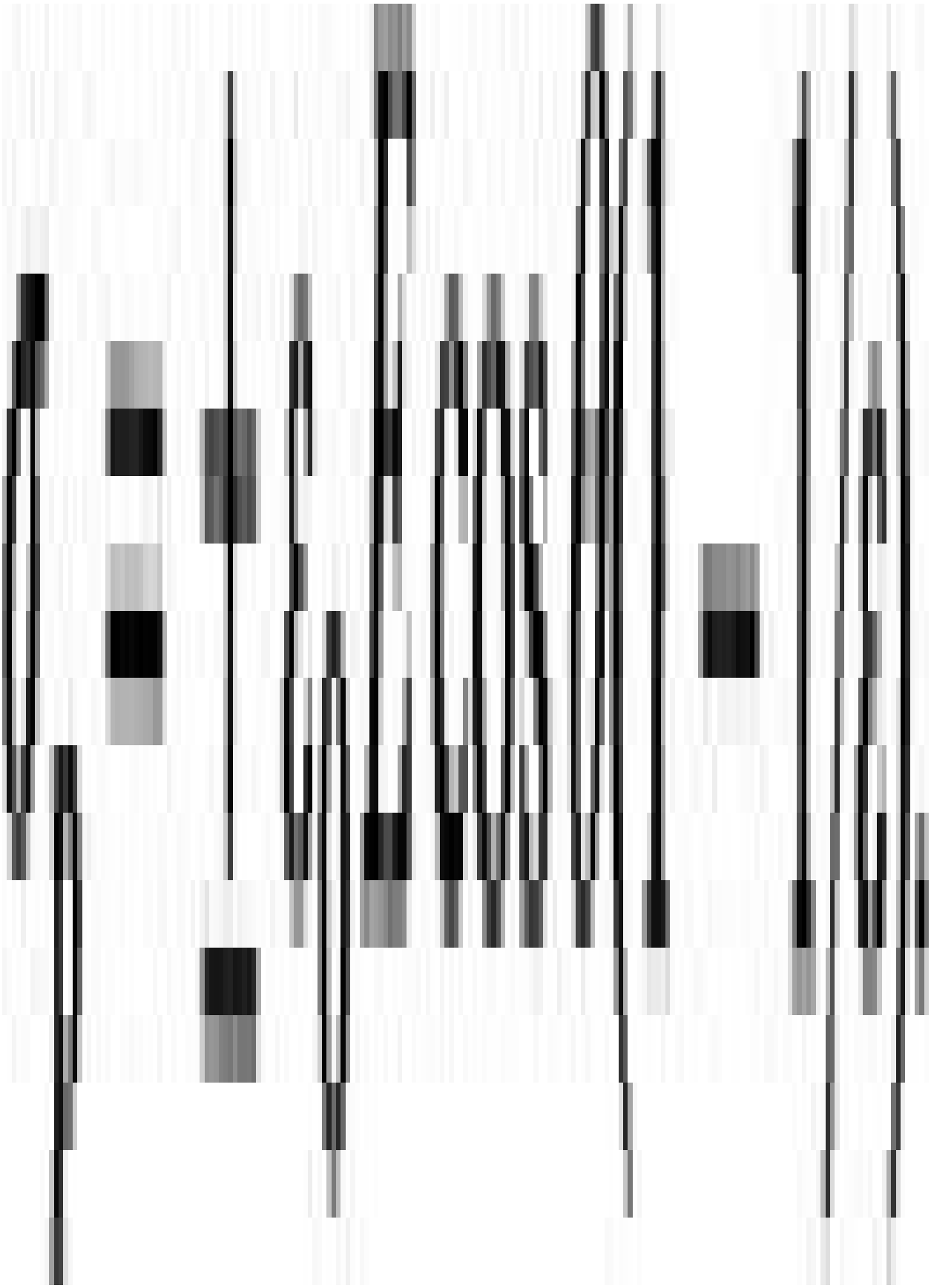
2.4 Elocal/E 6.6. Notice that the Clausius–Mossotti formula (Equation (2.16)) is hopelessly inaccurate for a polar liquid like water. In water the dipole moments of neighbouring molecules interact strongly with one another, and the problem of making a theoretical estimate of the average local field becomes very difficult.

2.5 The force between two dipoles varies as $1/r^4$. The potential energy of the two dipoles is (− 0.14) eV.

2.7 At the surfaces ± a there is a polarization charge density

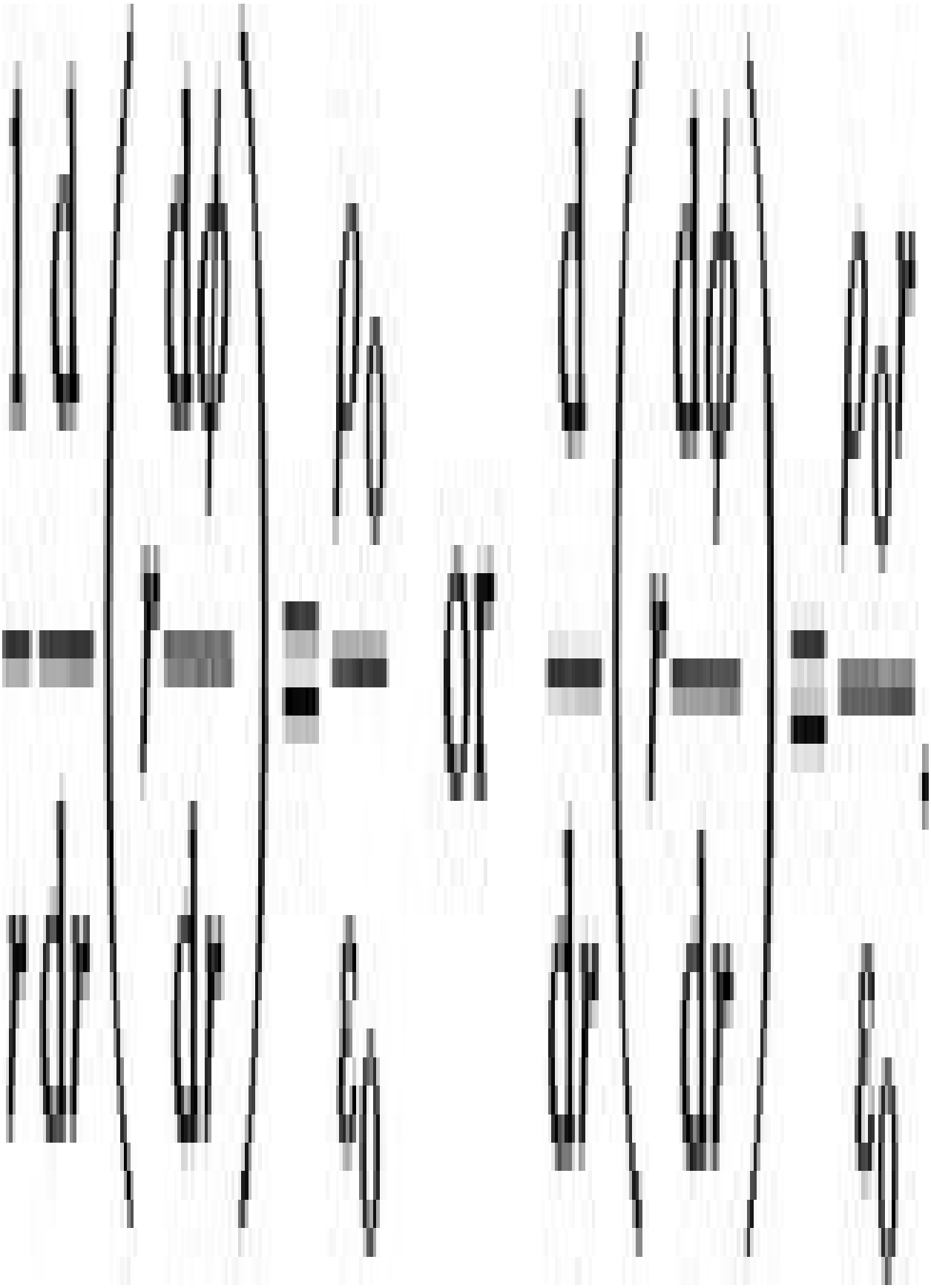


2.8 The density of polarization charge on the surface of the slab is



PROBLEMS 3

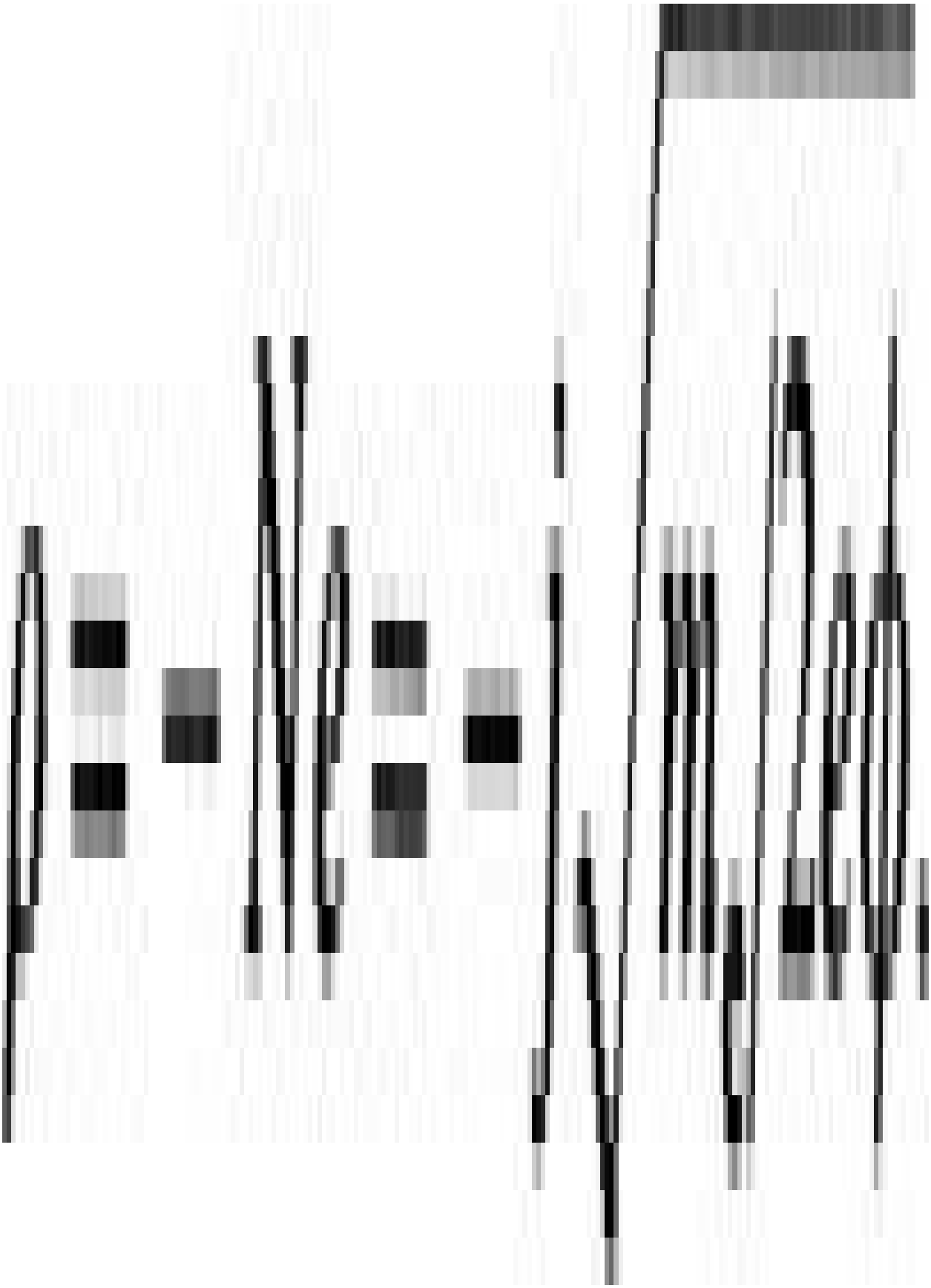
3.1 Poisson’s equation for the negative charge density $-\rho_0$ is



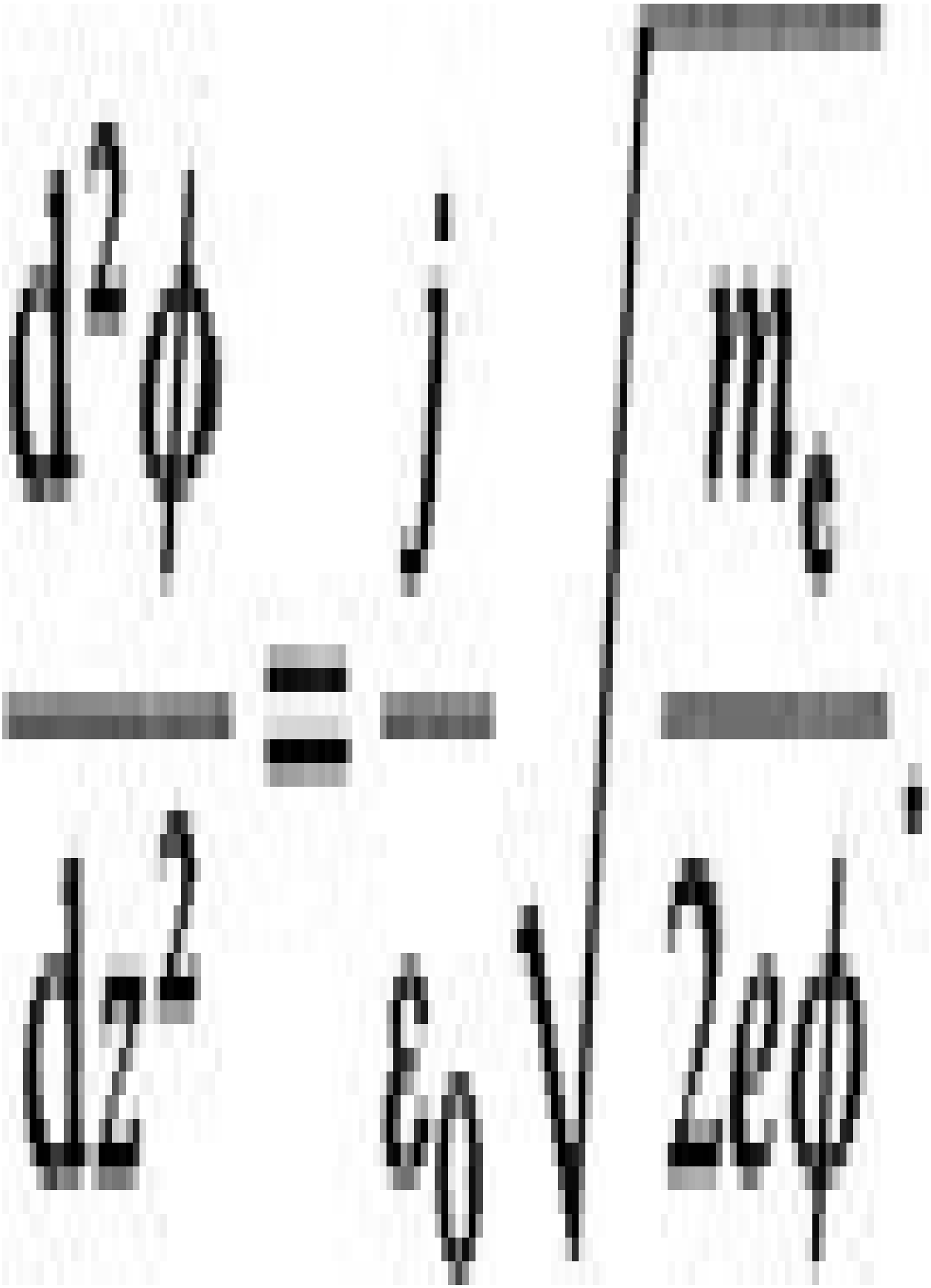
The solution for $r < R$ is $E0 = -d\phi/dr = -\rho 0r/2\varepsilon 0$. The constant of integration is zero because the field must be zero in the centre of the beam. For $R > r$ the charge density is zero, and $rd\phi/dr = c$, a constant. Since $E = -\rho 0r/2\varepsilon 0$ at $r = R$, $c = -\rho 0R^2/2\varepsilon 0$, and $E = -\rho 0R^2/2\varepsilon 0r$. As this problem has a simple symmetry, it can also be solved using the integral form of Gauss' law.

3.2 (i) If the cathode is at $\phi = 0$, then at the point where the potential is ϕ the electron speed v satisfies the equation $mev^2 = e\phi$.

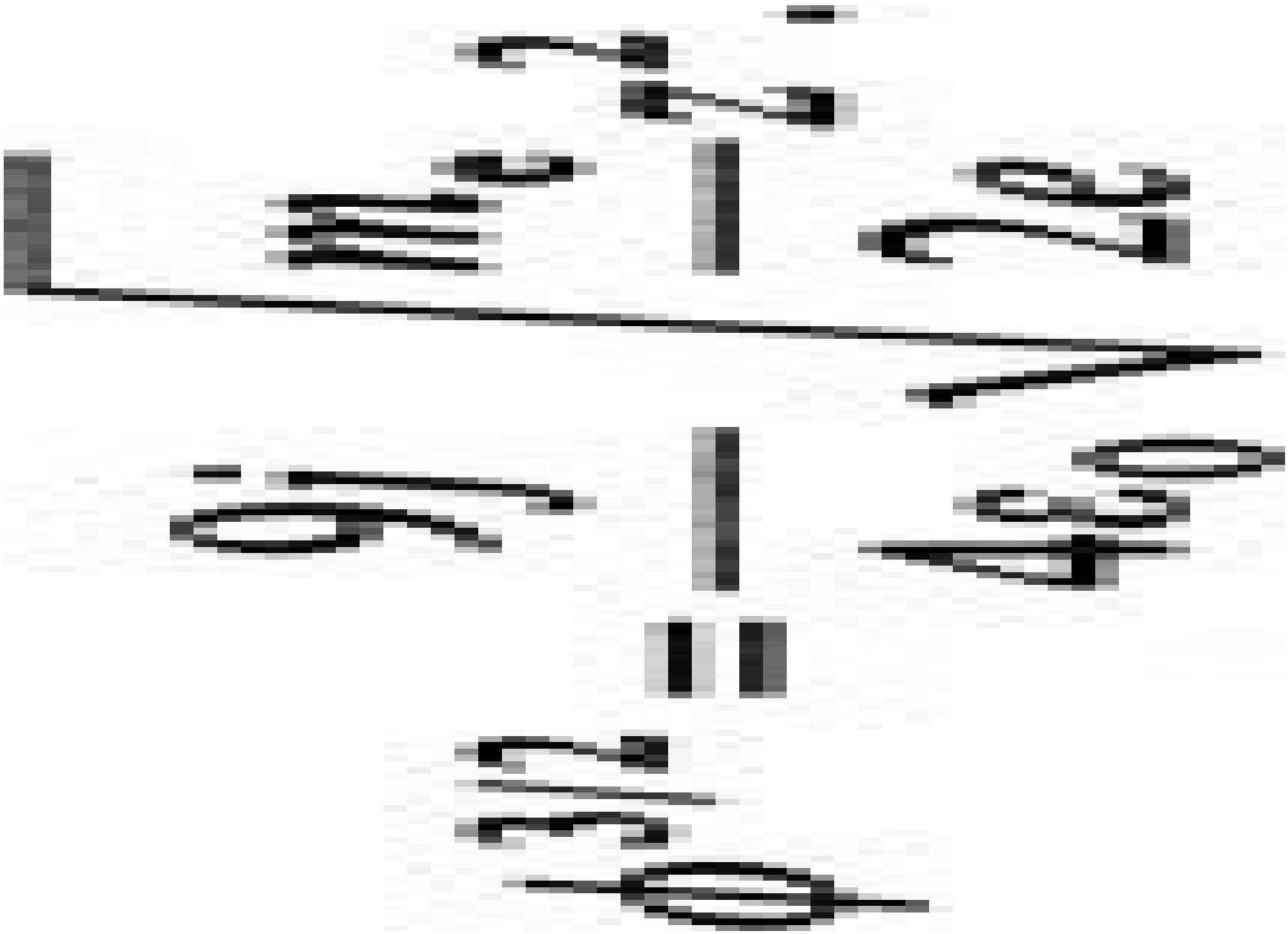
(ii) The magnitude of the current density is $j = Nev$, and



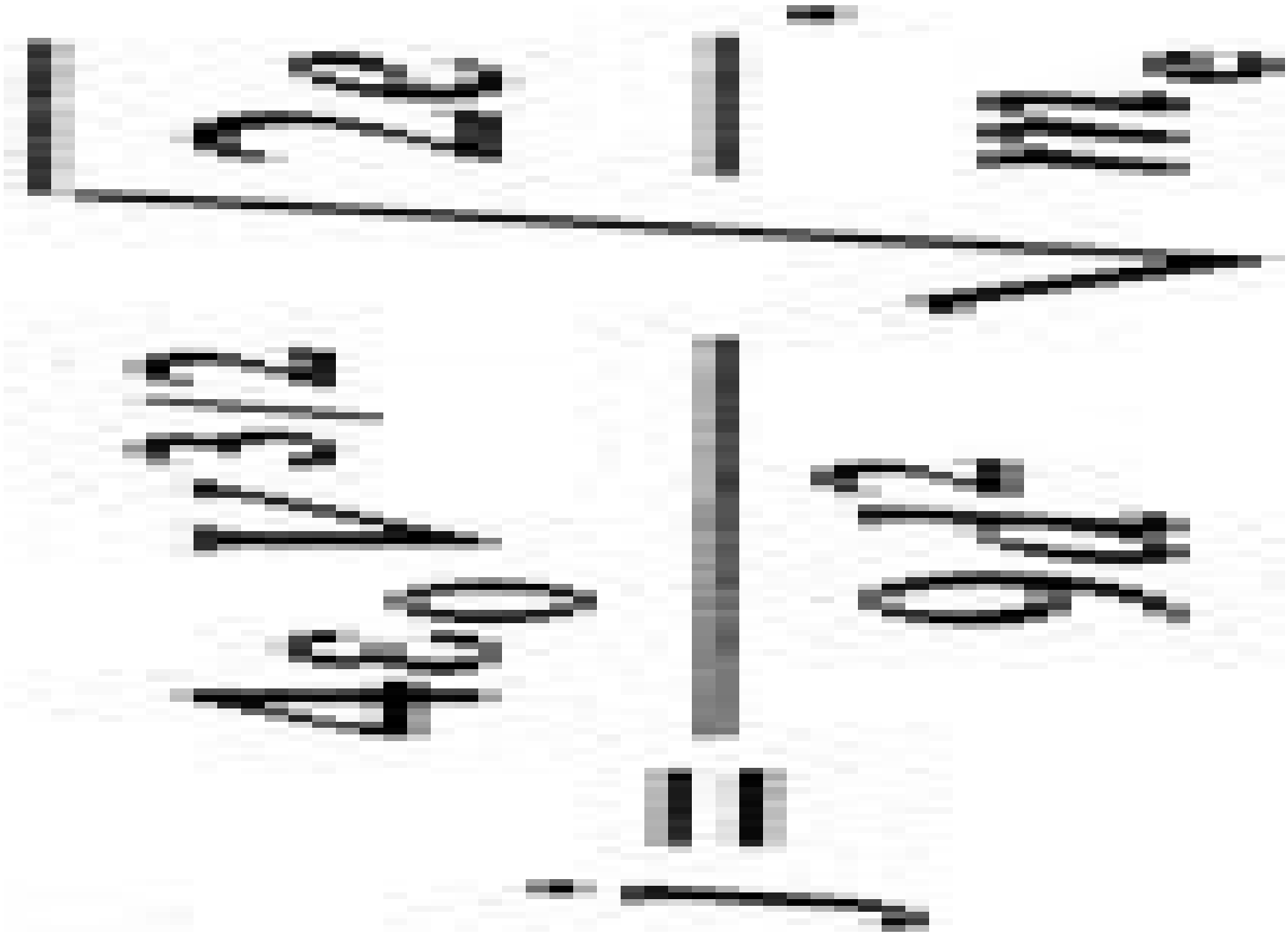
(iii) Poisson’s equation is



The solution which satisfies the boundary conditions is



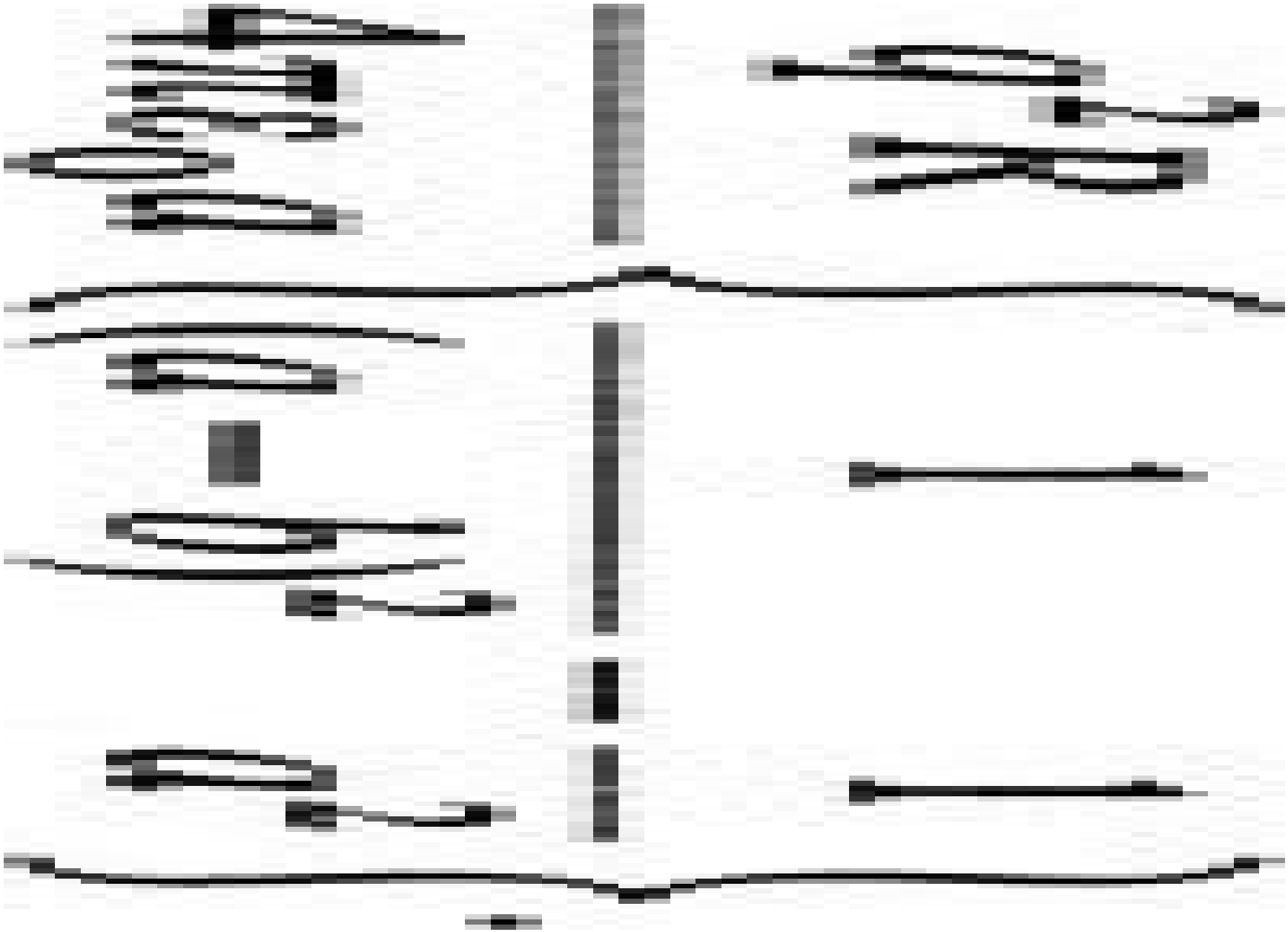
Thus when a potential difference V is maintained across plates a distance d apart, the current density is



3.3 The electrostatic force attracting the proton towards the image charge (−e) 2 cm away is 36 times the gravitational force acting on it.

3.4 5.3 pF/m.

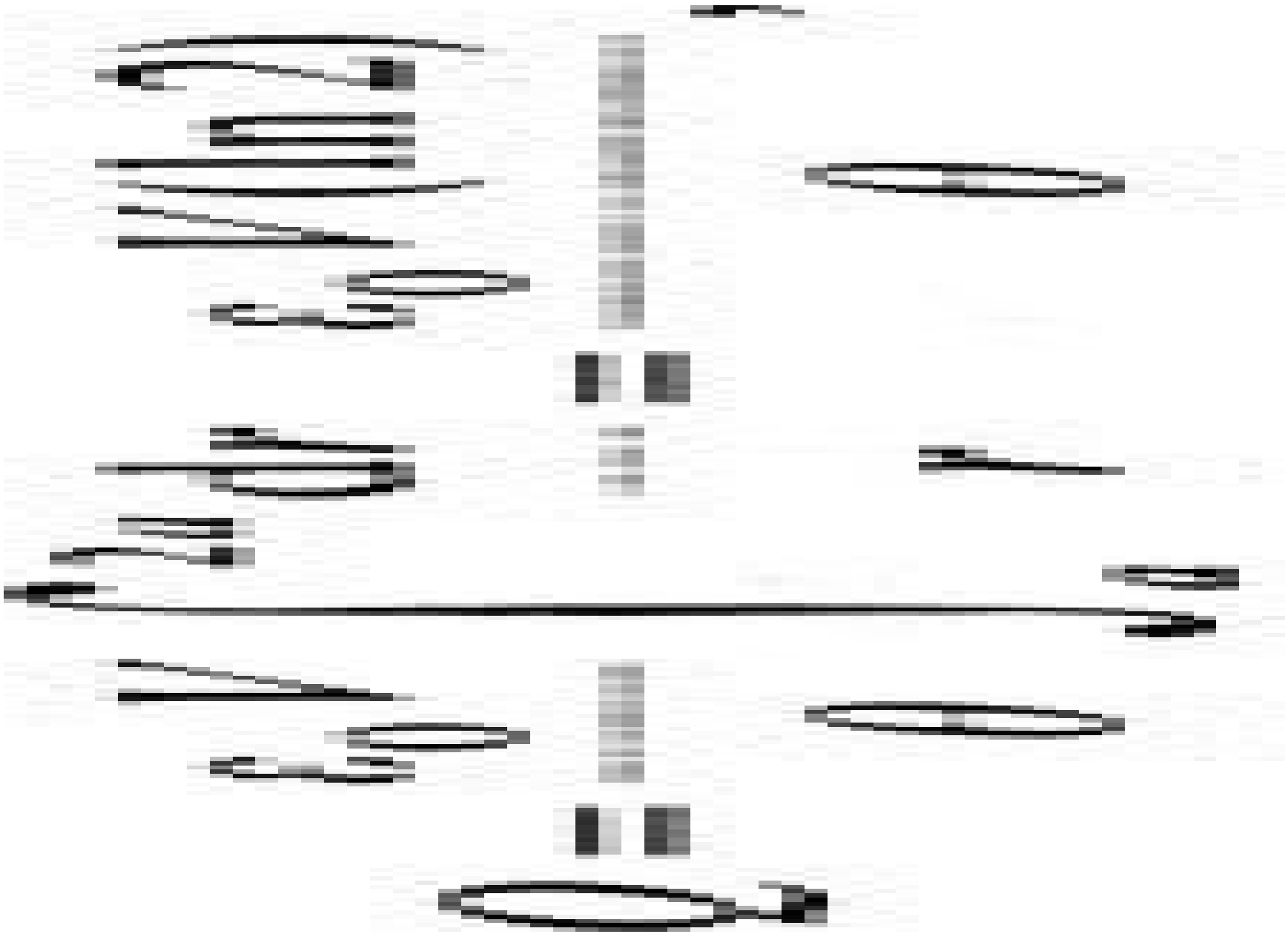
3.5 If the sphere carries no net charge it can be represented by q′ and an additional charge of magnitude + qR/a situated at the centre of the sphere. The force on q is



3.6 In the previous problem the charge q and its image q′ have different magnitudes. It follows that two equal and opposite image charges cannot exactly represent the two charged spheres. However, since their separation is a good deal larger than their radii, it is not too bad an approximation to represent them by point charges ± Q located at their centres. In this approximation the potential difference between the spheres along the line of their centres is 85 kV when the field at the surface is on the point of breakdown. A better approximation is to represent each sphere by a charge ± 0.2Q at a distance 0.4 cm from the centre (this is the image of Q located at the centre of the other sphere) and an additional ± 0.8Q at the centre to preserve the correct magnitude of the charge. This leads to an estimate of 81 kV for the breakdown voltage. A further approximation can be made by taking the images of both these charges, and continuing the process the spheres can be exactly represented by an infinite series of image charges. Only a few terms in the series are needed to give a good answer, since we have seen that the first and second approximations differ by only about 3%.

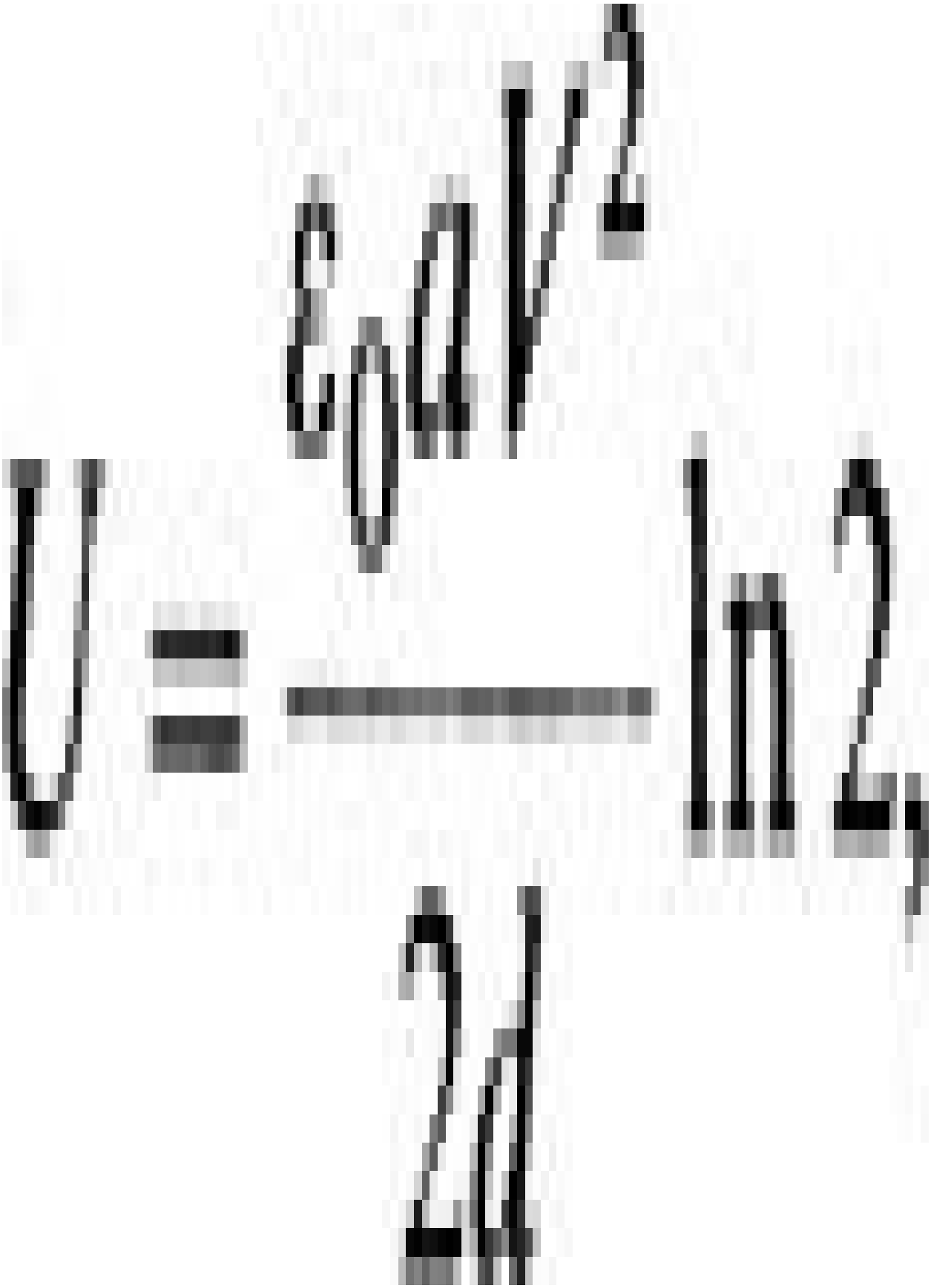
3.7 The induced charge density on each plate is $-\sigma/2$.

3.9 Within the wedge capacitor, the length of a circular arc at a radius r is $r\theta$, and the electric field is $E = V/r\theta$. The field lines terminate perpendicularly on the plate, satisfying the boundary conditions for E, and inducing a surface charge density $\sigma = \varepsilon_0 V/r\theta$. The charge on a unit length of the capacitor is

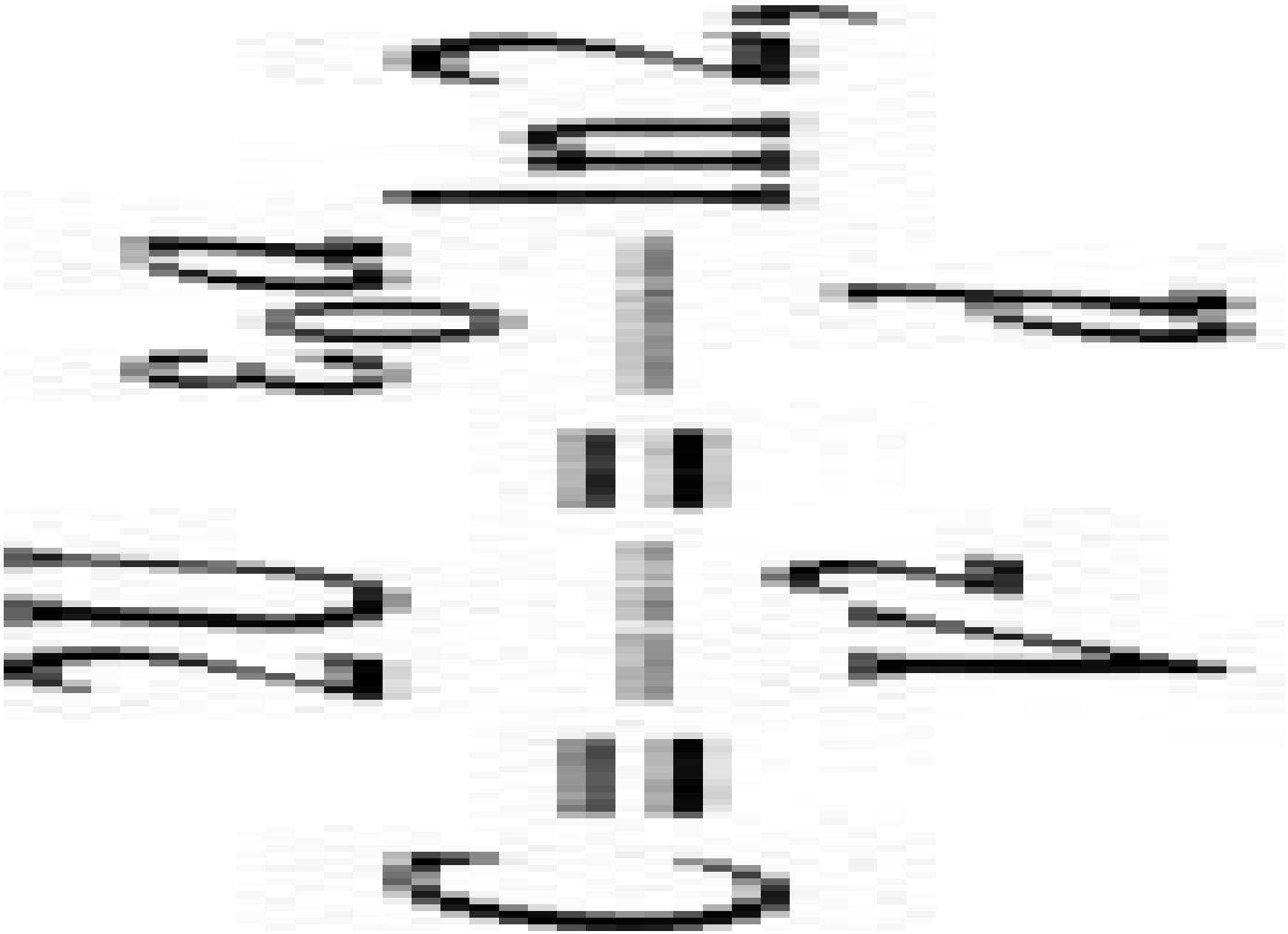


and the capacitance per unit length is $\varepsilon_0(\ln 2)/\theta$.

3.10 Assuming that the field lines are straight, the total energy stored between the plates per unit length of the capacitor is



and hence



a good approximation to the value ε0 In 2/θ found in the last problem. An approximate field pattern always leads to a result for the stored energy a little greater than the exact value; the pattern of an electrostatic field is the one which minimizes the stored energy.

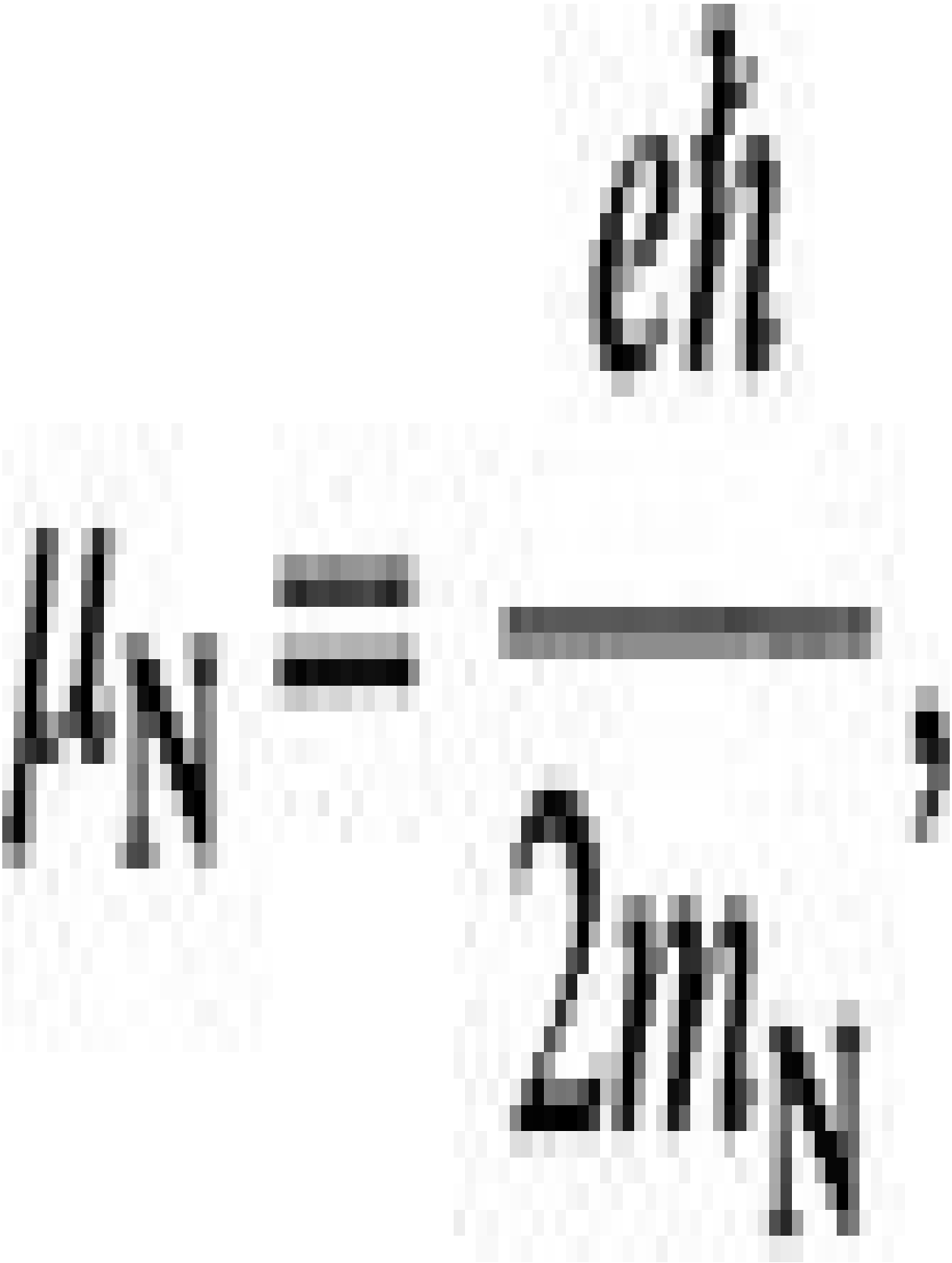
PROBLEMS 4

4.1 After a collision with a molecule in the gas the momentum of an ion is, on average, zero. Hence the mean speed is equal to one half of the speed gained between collisions, which can be worked out by calculating the acceleration of the ions due to the applied electric field. Remember that the thermal speeds are much greater than the speeds acquired by acceleration. The mean drift speed is 0.5 m s−1. The current is given by Equation (4.3) to be 2.1 × 10−8 A.

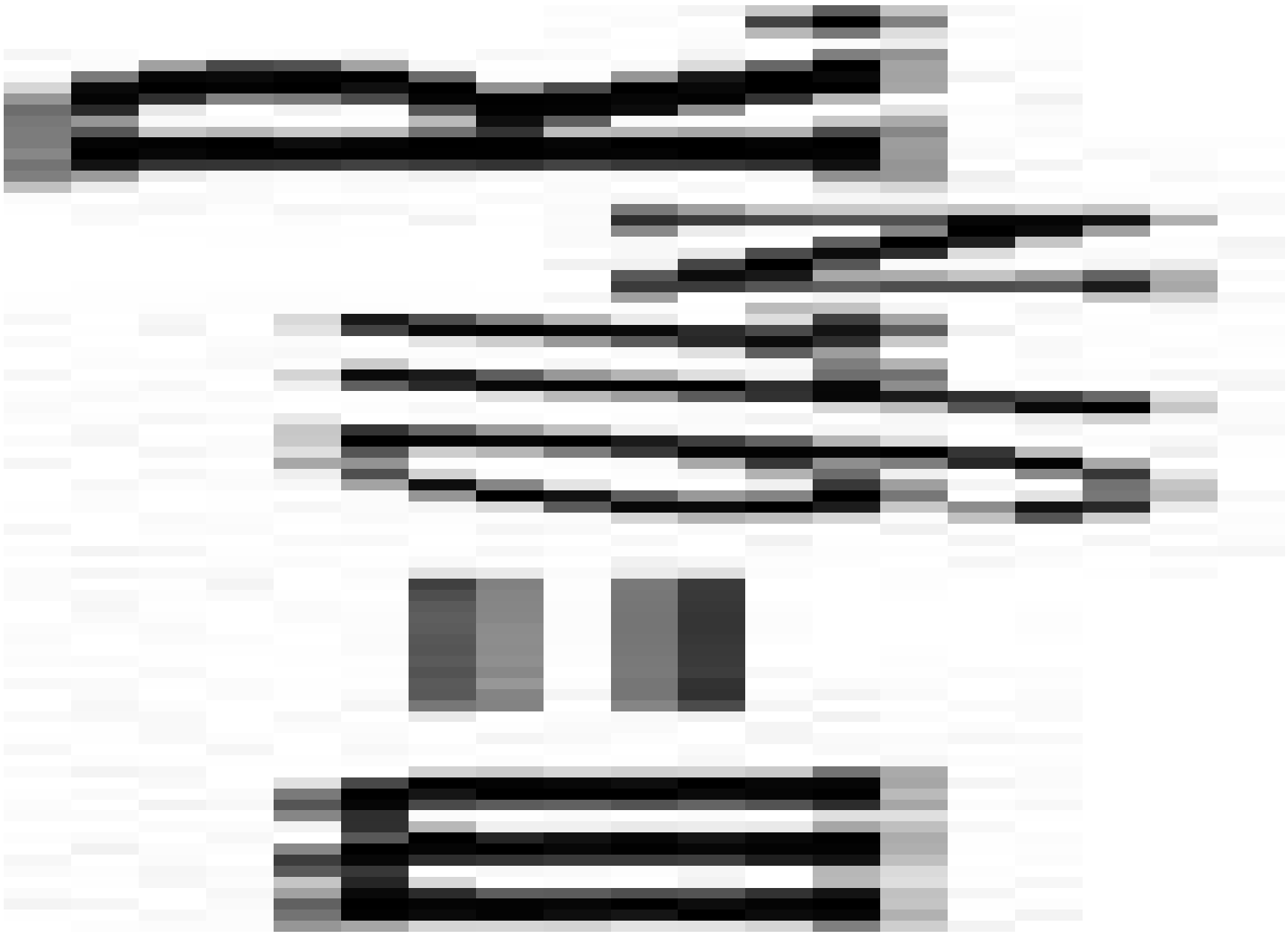
4.2 See Problem (1.6). The resistance between each sphere and Earth is ρ/2πa. The resistance between the spheres equals ρ/πa.

4.3 The magnetic dipole moment equals LQ/2M. The nucleus of an atom which has atomic number Z is quite well represented by a uniformly charged sphere carrying total charge + Ze. If it contains A nucleons (protons or neutrons) its mass is ~ mNA, where mN is the nucleon mass (nearly the same for protons and neutrons). The angular momentum L of a rotating nucleus = R where R is an integer and is Planck's constant divided by 2π (see section 4.4.3). Hence the magnetic dipole moment m of the rotating nucleus is equal to R Ze/2mNA. This may be rewritten as

where μN, the nuclear magneton, is given by



and R is a vector in the direction of the angular momentum. The factor Z/A is called the rotational gyromagnetic factor, and the magnetic moment may be written as



4.4 The maximum field gradient in the x − y plane is 10.4 tesla m−1. The force on a moving atom which is neutral, but which possesses a magnetic dipole moment, will be a maximum if the atom is moving in the x-direction at that value of the coordinate y where the field gradient is a maximum. Use is made of this in the Stern–Gerlach experiment.

4.5 The field can most easily be obtained by solving the equation curl B = μ0j inside and outside the sheet, and noting that the field must be independent of the coordinates x and z. The field then has an x-component only, which varies with coordinate y as shown in Figure D.l. The force on an electron moving near the speed of light perpendicular to a field π × 10−2 tesla is 1.51 × 10−12 newtons. In some circular machines that accelerate particles to high energies a large current is pulsed through a conducting sheet at the time a burst of beam comes near. The magnetic field produced deflects the charged particle beam out of its path in the accelerator and into experimental apparatus.

4.7 The principle of superposition is true for magnetic fields. The field of a cable with a hole drilled in it and carrying a uniform current density j is the same as the field of a solid cable carrying that current density plus the field of a current density –j in the opposite direction occupying the space where the hole was.

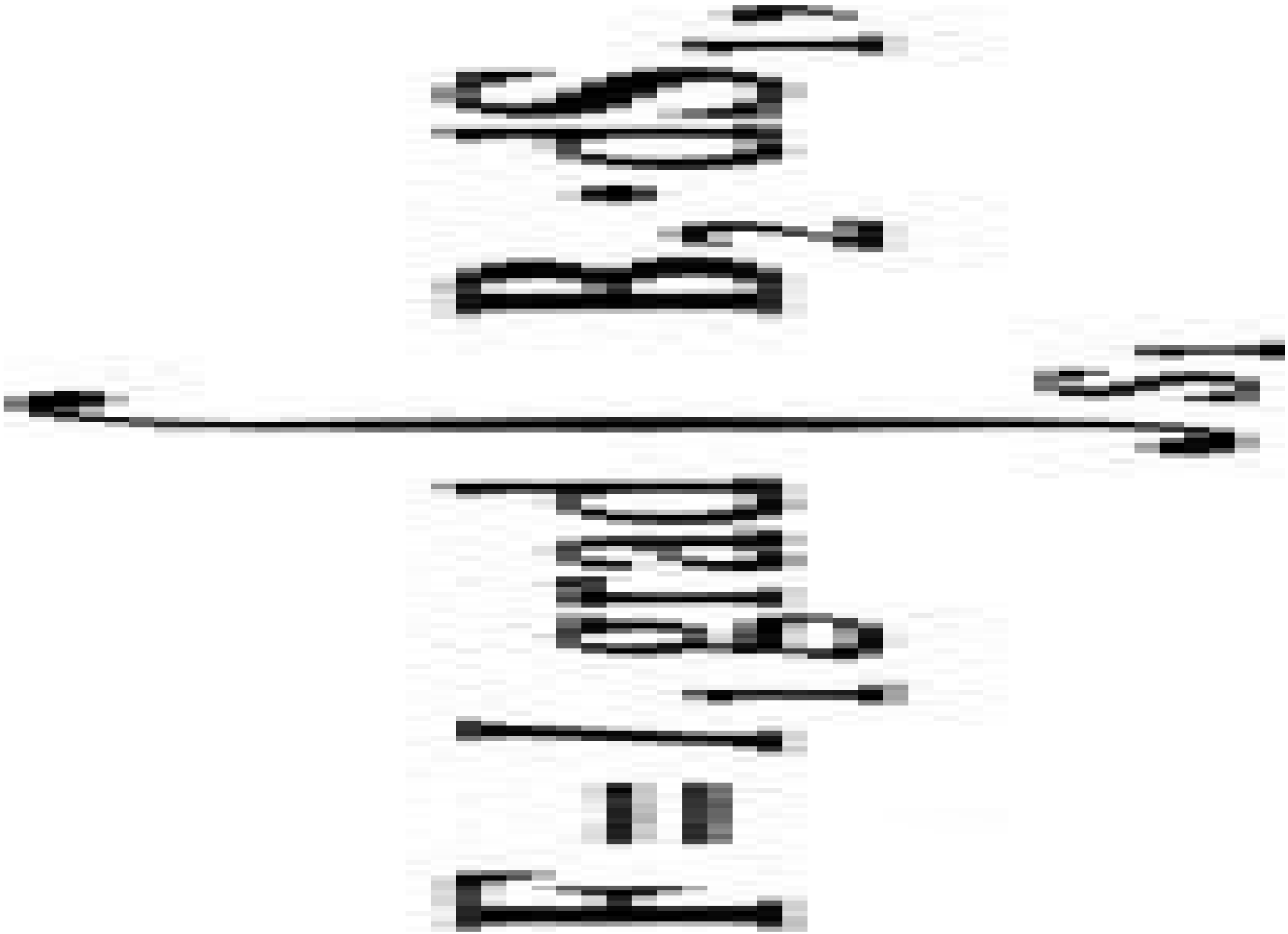
Answer: (i) π × 10−6 teslas (ii) π × 10−4 teslas.

4.9 Assume that the field from the large coil is constant over the small coil. The force is given by Equation (4.50). Answer: 1.4 × 10−4 N.

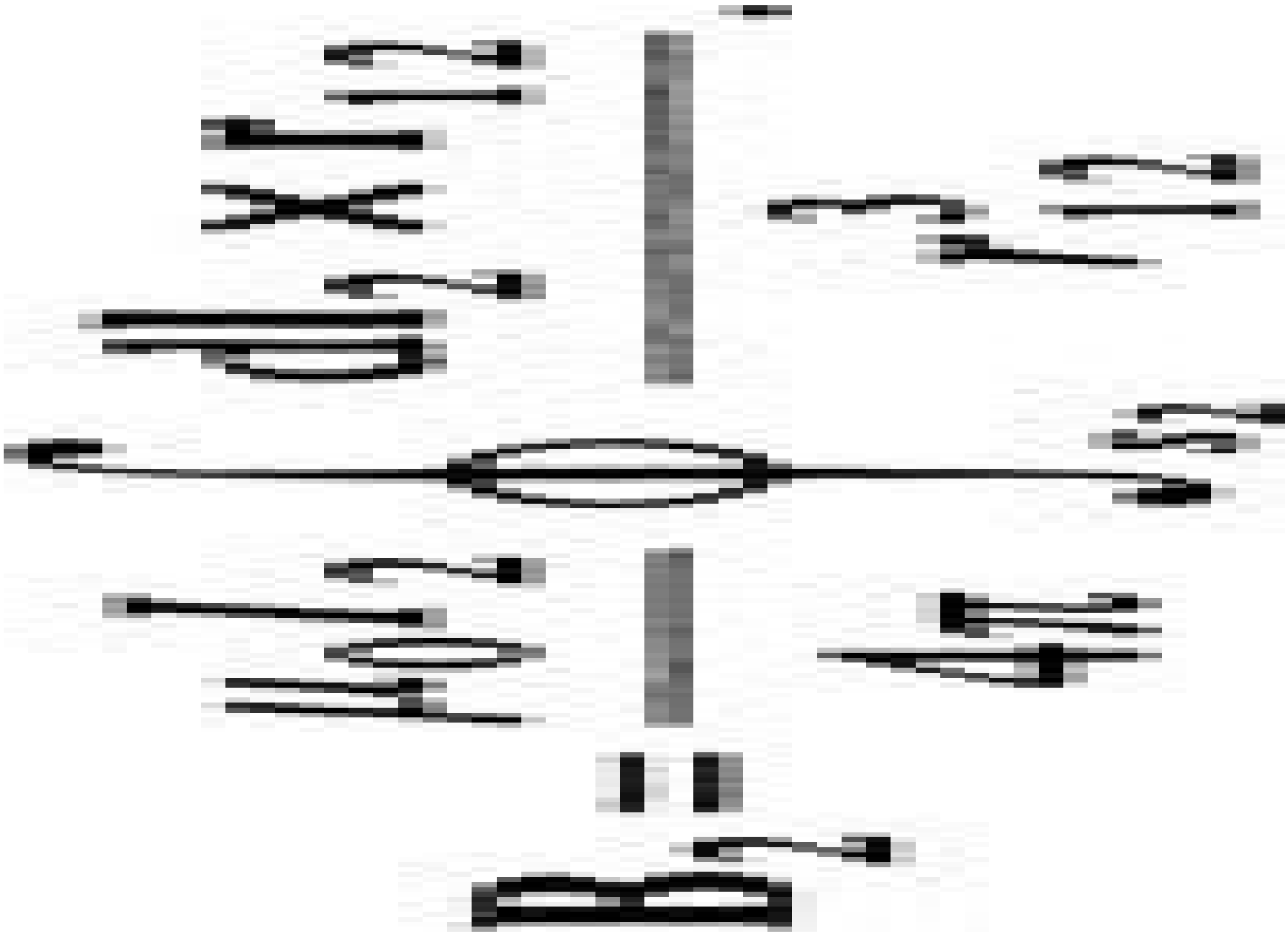
4.12 The motion of the electron parallel to the magnetic field is unaffected by the magnetic force. The motion in the plane perpendicular to the field is circular with radius given by Equation (4.58). The electrons thus follow the magnetic field lines, spiralling about them in a helix. The number of revolutions made is equal to 8.8×10^{10}.

4.14 5.562 × 10−23 N.

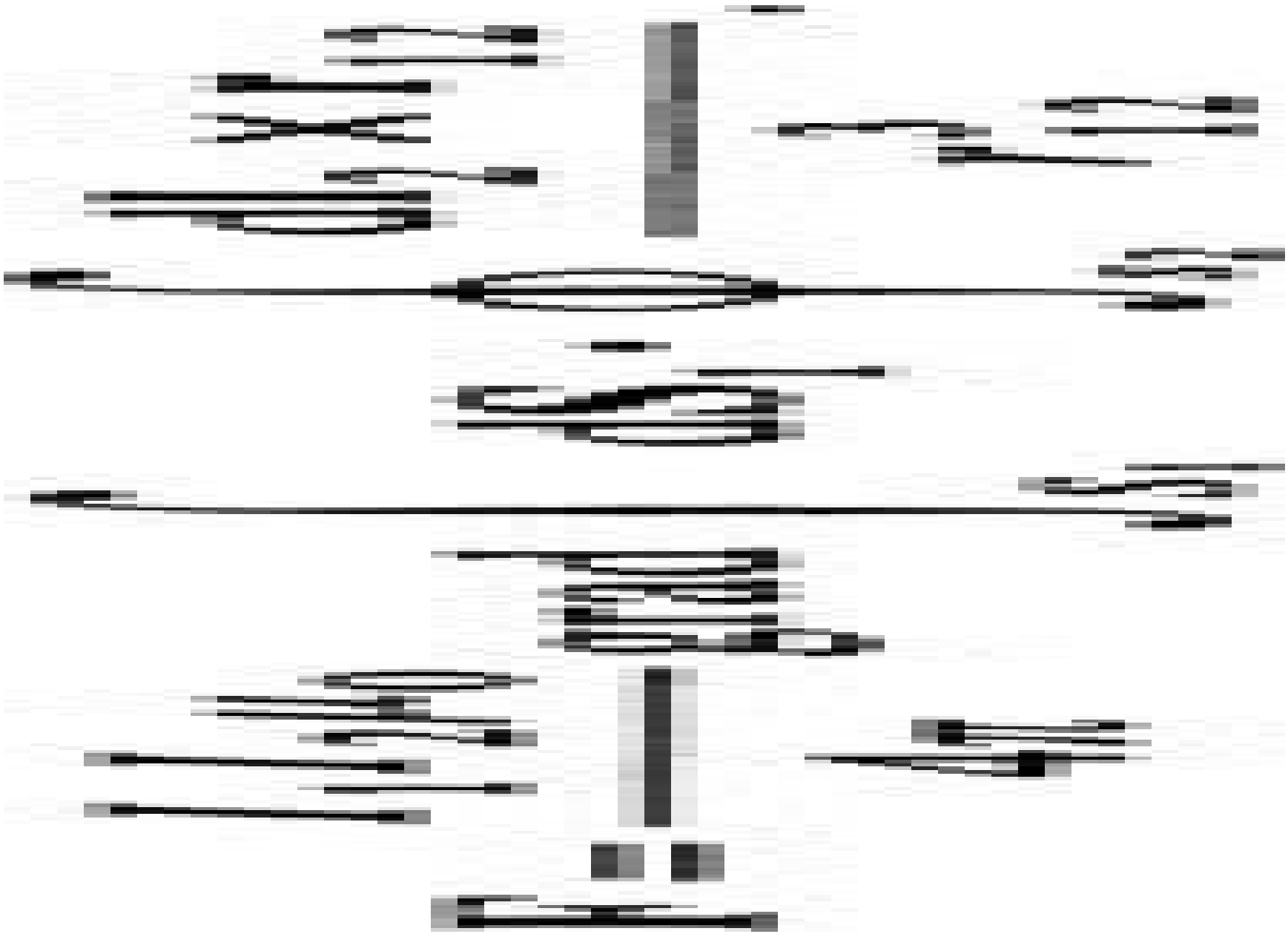
4.16 The Biot–Savart law and Equation (4.53) give



where



Hence

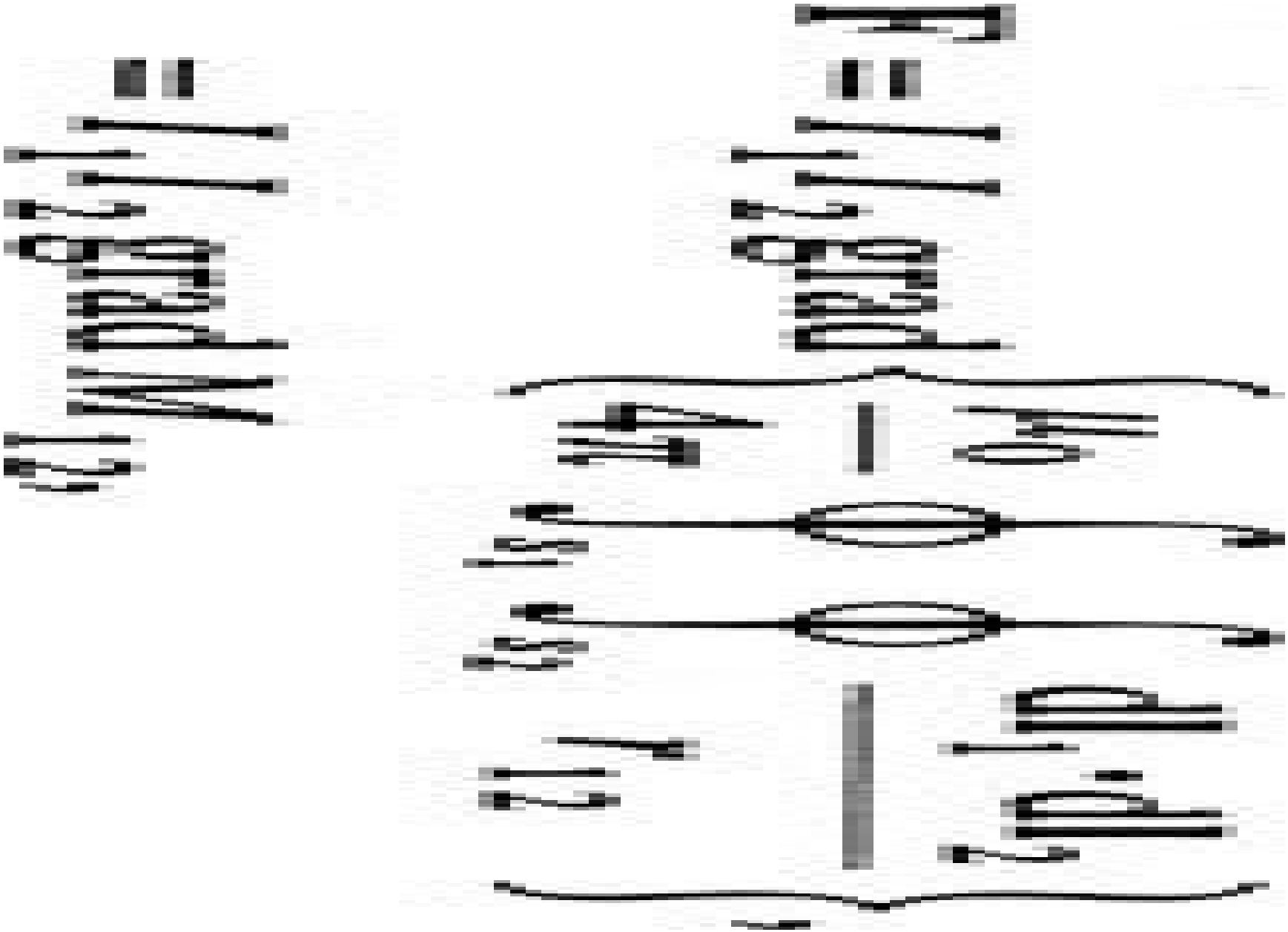


The Biot–Savart law and Equation (4.50) give

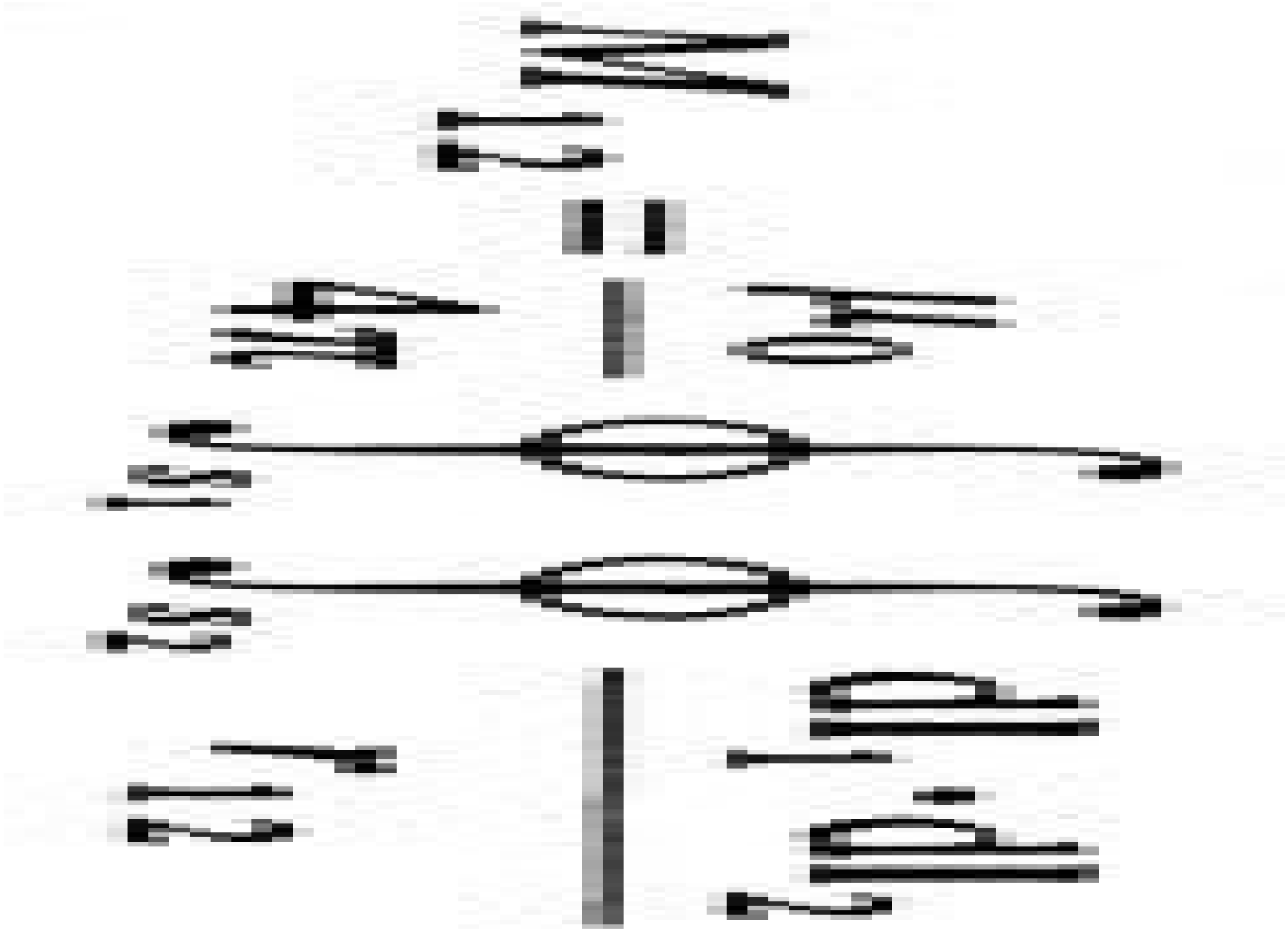
$$\mathbf{F} = I_1 \oint_{S_1} d\mathbf{l}_1 \times \mathbf{B}_2$$

$$= I_1 \oint_{S_1} d\mathbf{l}_1 \times \oint_{S_2} \frac{\mu_0 I_2 \, d\mathbf{l}_2 \times \mathbf{r}_{12}}{4\pi \, r_{12}^3}$$

Both can be shown to reduce the expression



where

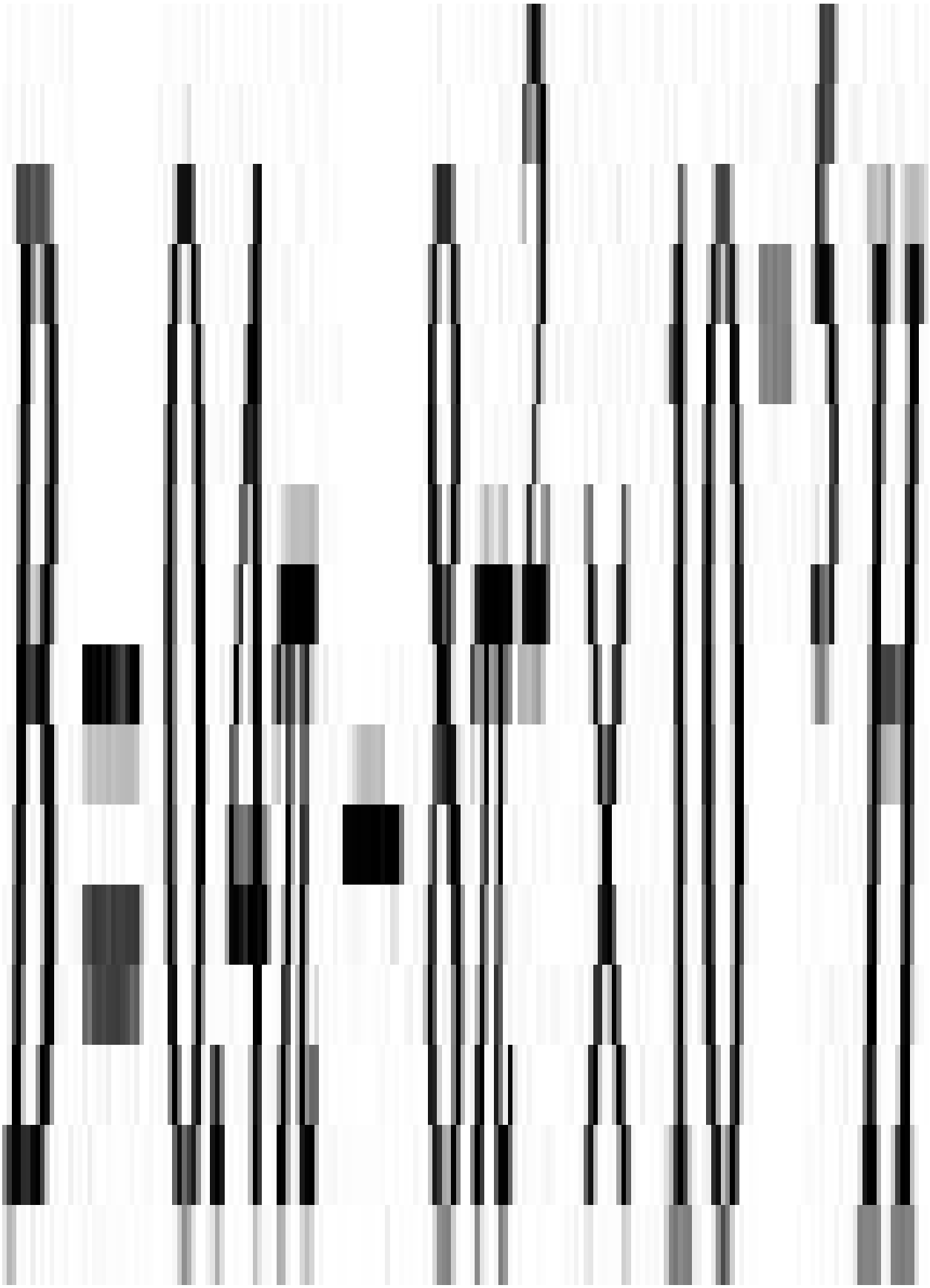


is the mutual inductance between the coils (see section 6.2.2) and the above formula for M12 is called Neumann’s formula.

PROBLEMS 5

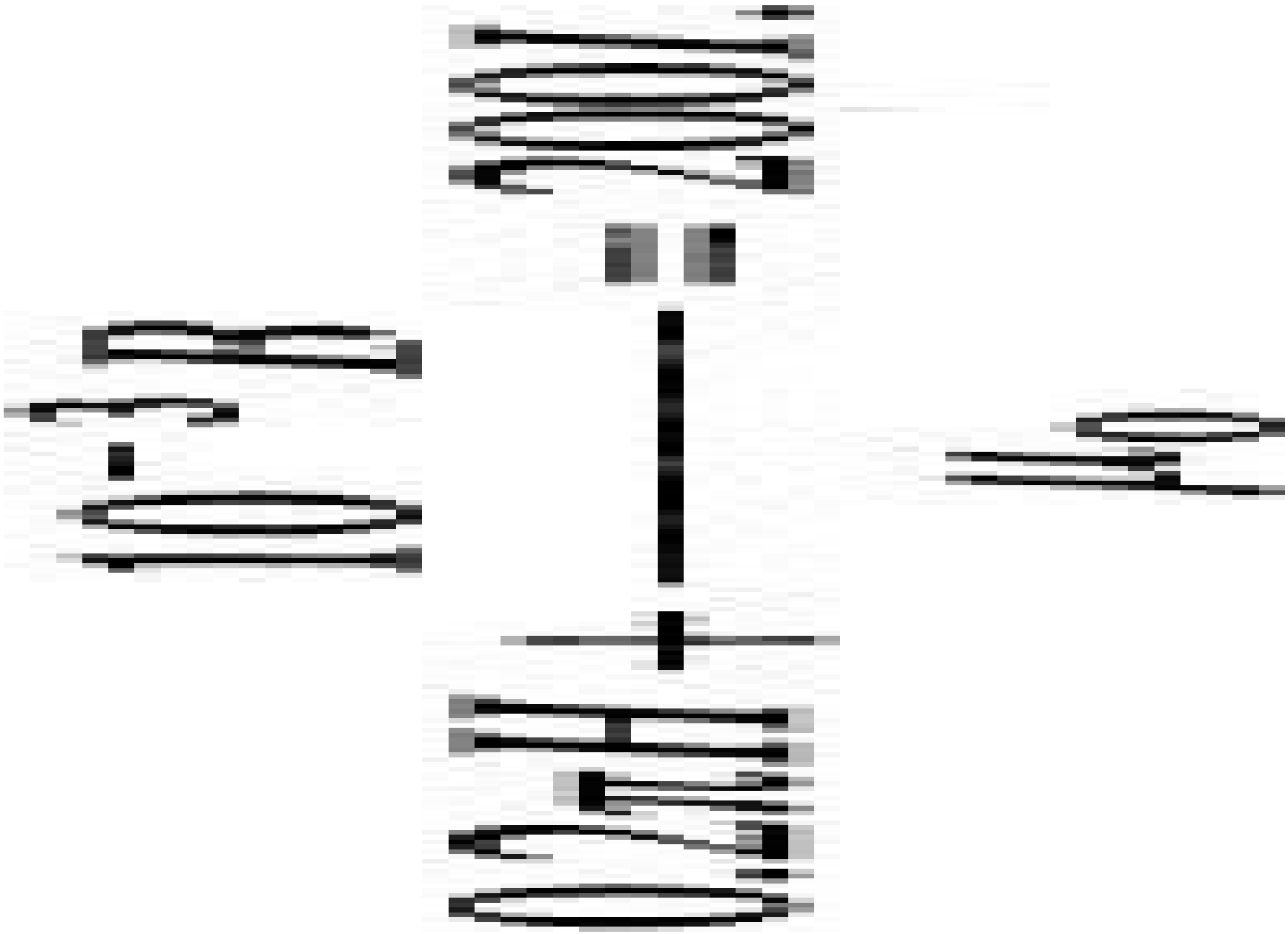
5.1 About 3×10^{-23} A m². The diamagnetic correction is about 15%.

5.3 The field is given by the intersection of the line



with the B–H curve plotted from the data given in the question. The intersection gives B = 1.0 T.

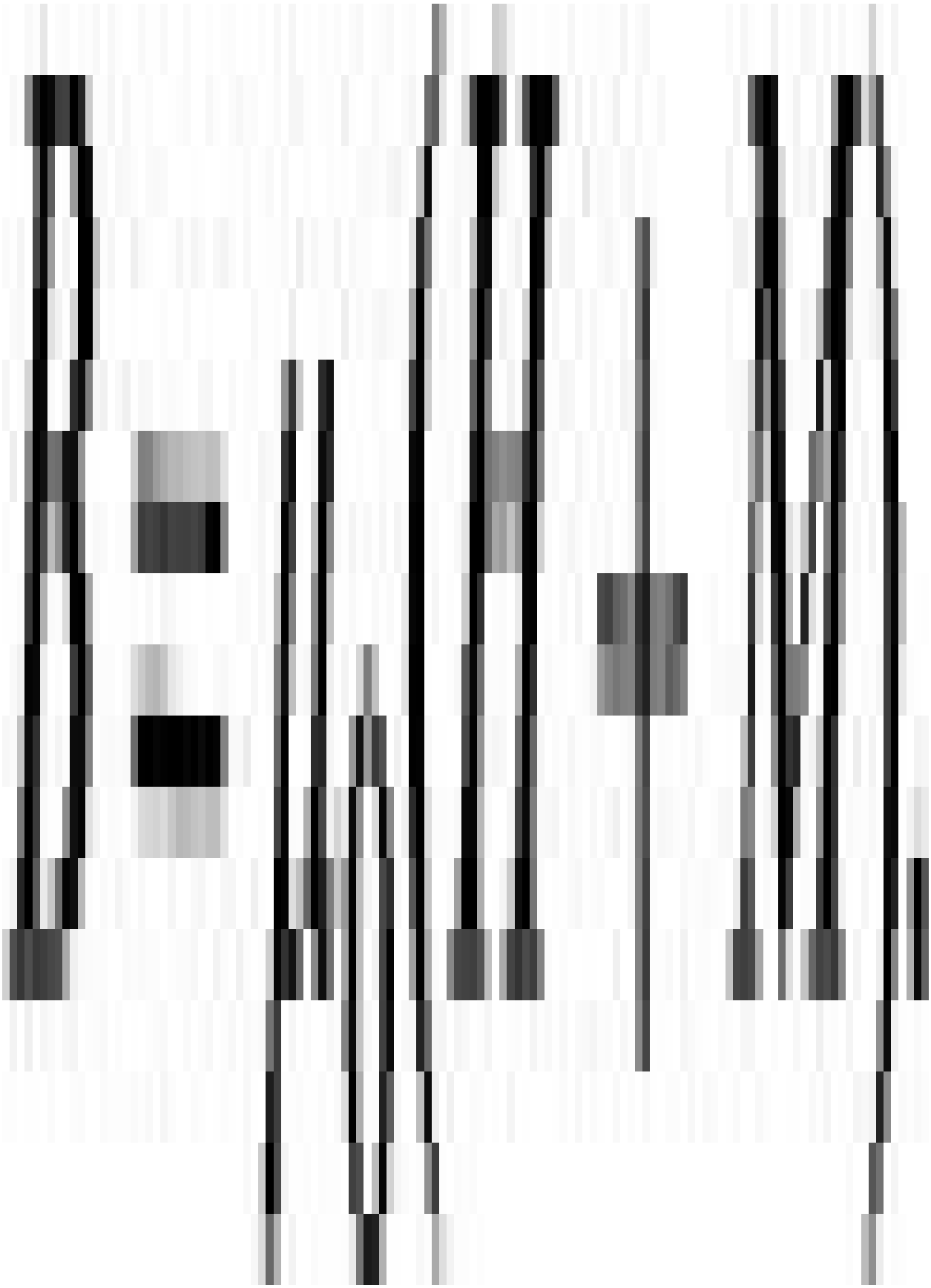
The current I required to produce a given field B is obtained from the equation (derived from Ampère’s integral law)



This gives I1 and I2. The power required is proportional to I^2, whence the power ratio is 0.21.

5.4 When one coil is disconnected, the yoke forms a complete magnetic circuit of low reluctance. Almost all the flux is trapped in the yoke, and the field in the gap falls to a low value.

5.6 The field B inside the magnet is related to the magnetic intensity H and the magnetization M by



The field H is in the opposite direction to B and M inside the magnet (in order to make the line integral H·dl zero over a closed loop through the magnet). The field H will be of the same order of magnitude inside the magnet to just outside; consequently it can be neglected compared with M and

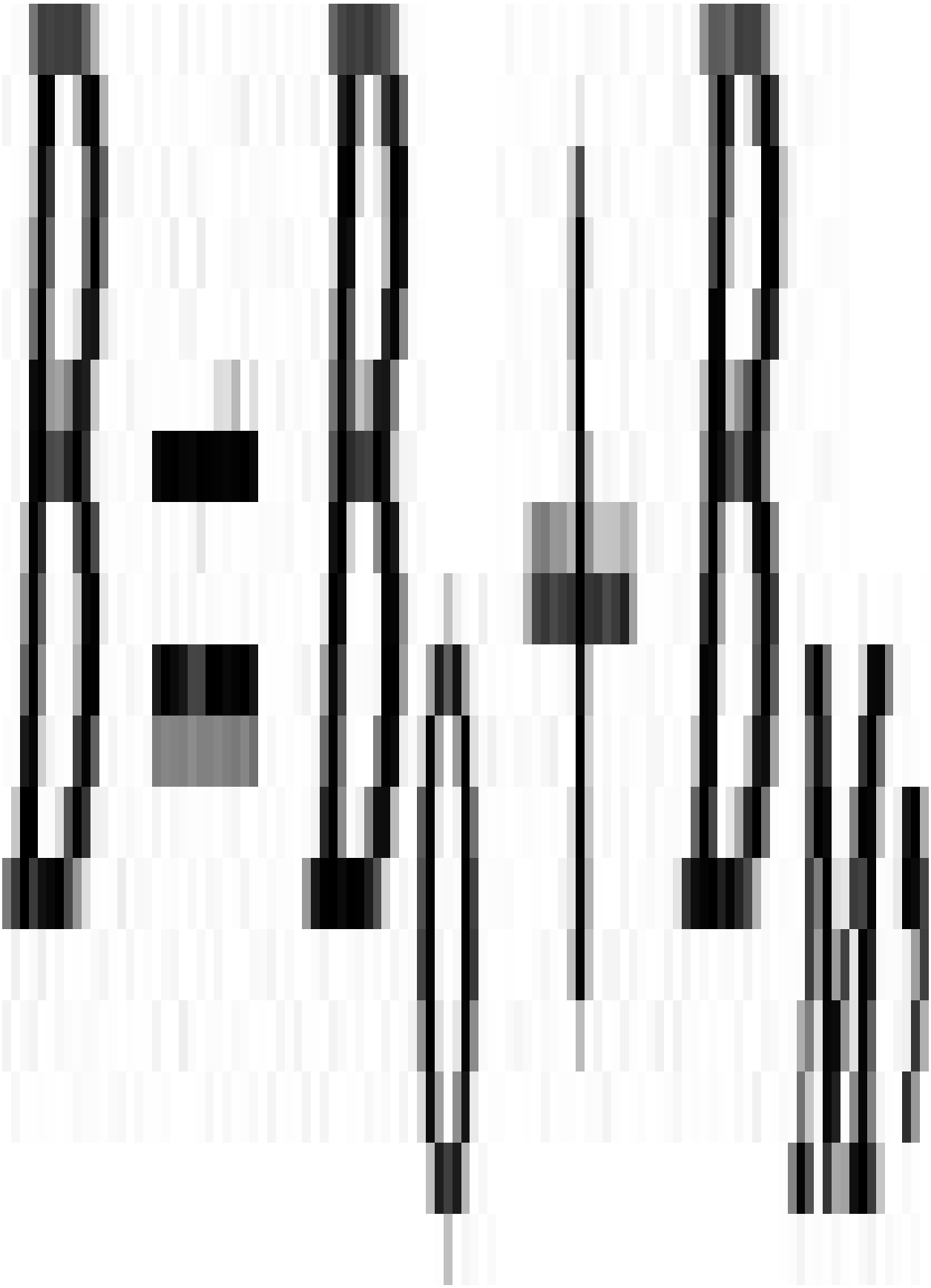
$$B \approx \mu_0 M.$$

$$M = \frac{\text{moment}}{\text{volume}} = \frac{75 \times 4}{\pi \times (0.05)^2 \times 0.1}$$

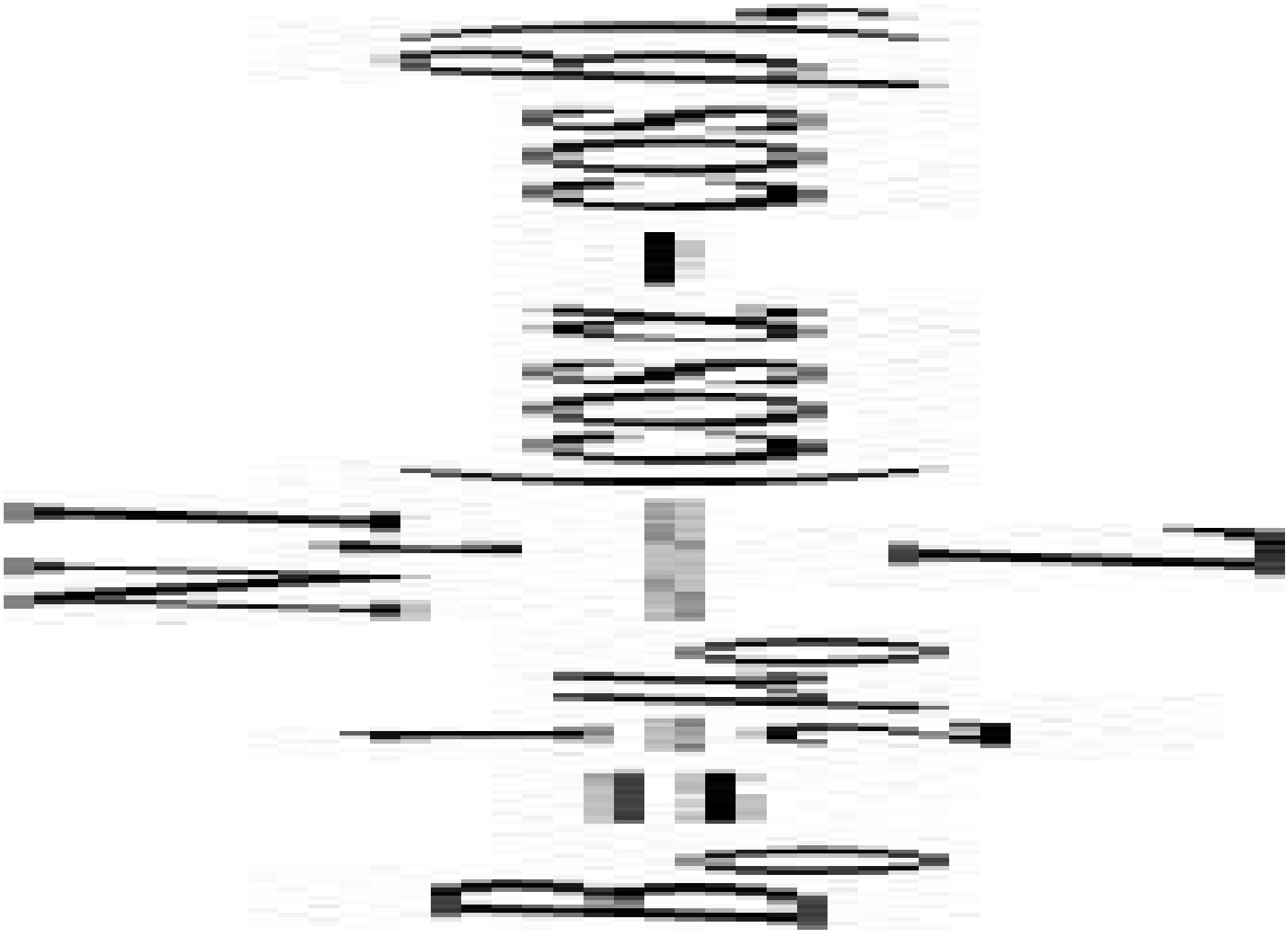
$$= 1.2 \times 10^6 / \pi \, \text{Am}^{-1}.$$

Hence the field B is about 0.48 T.

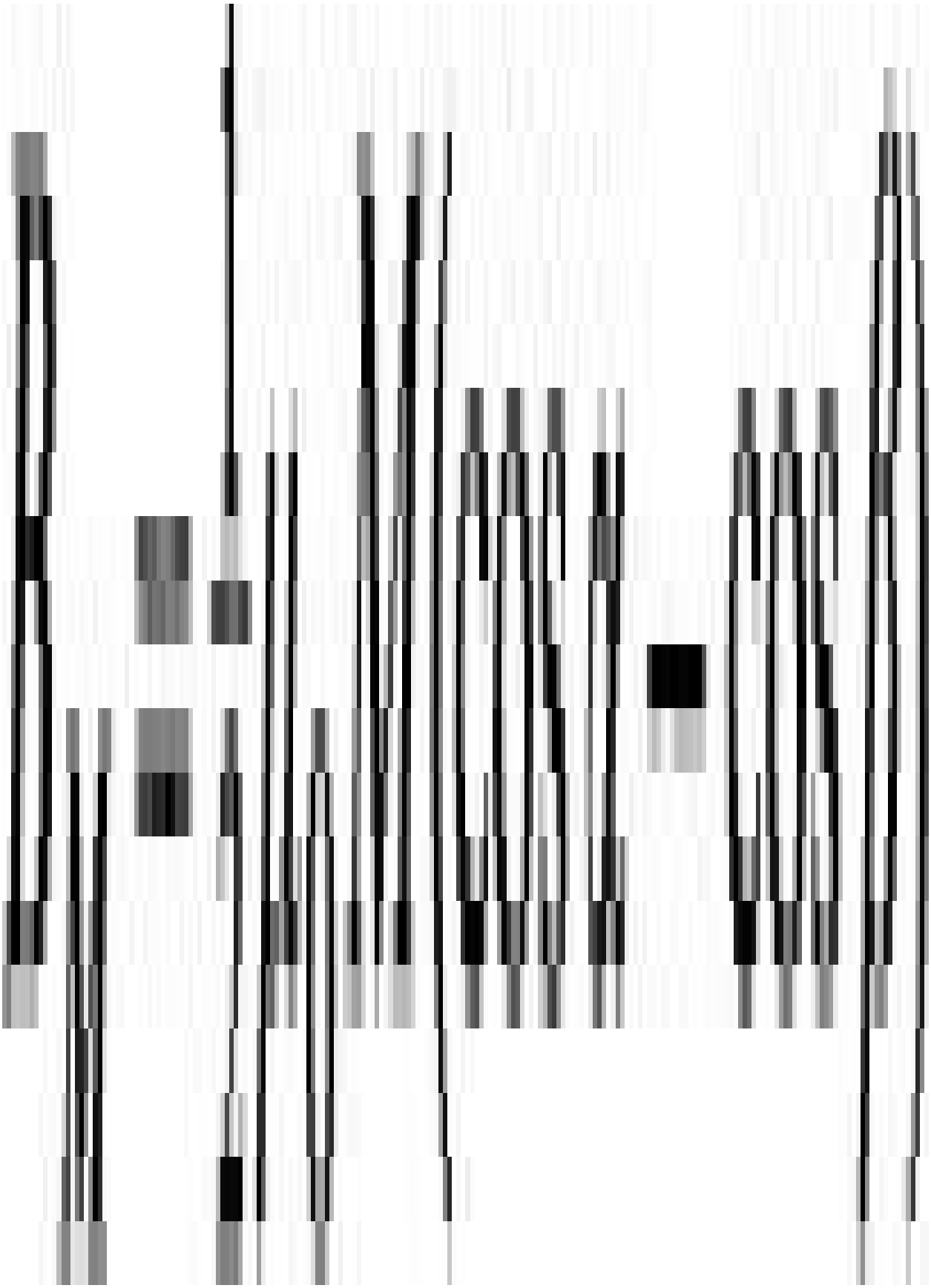
5.8 The total field at the point P is



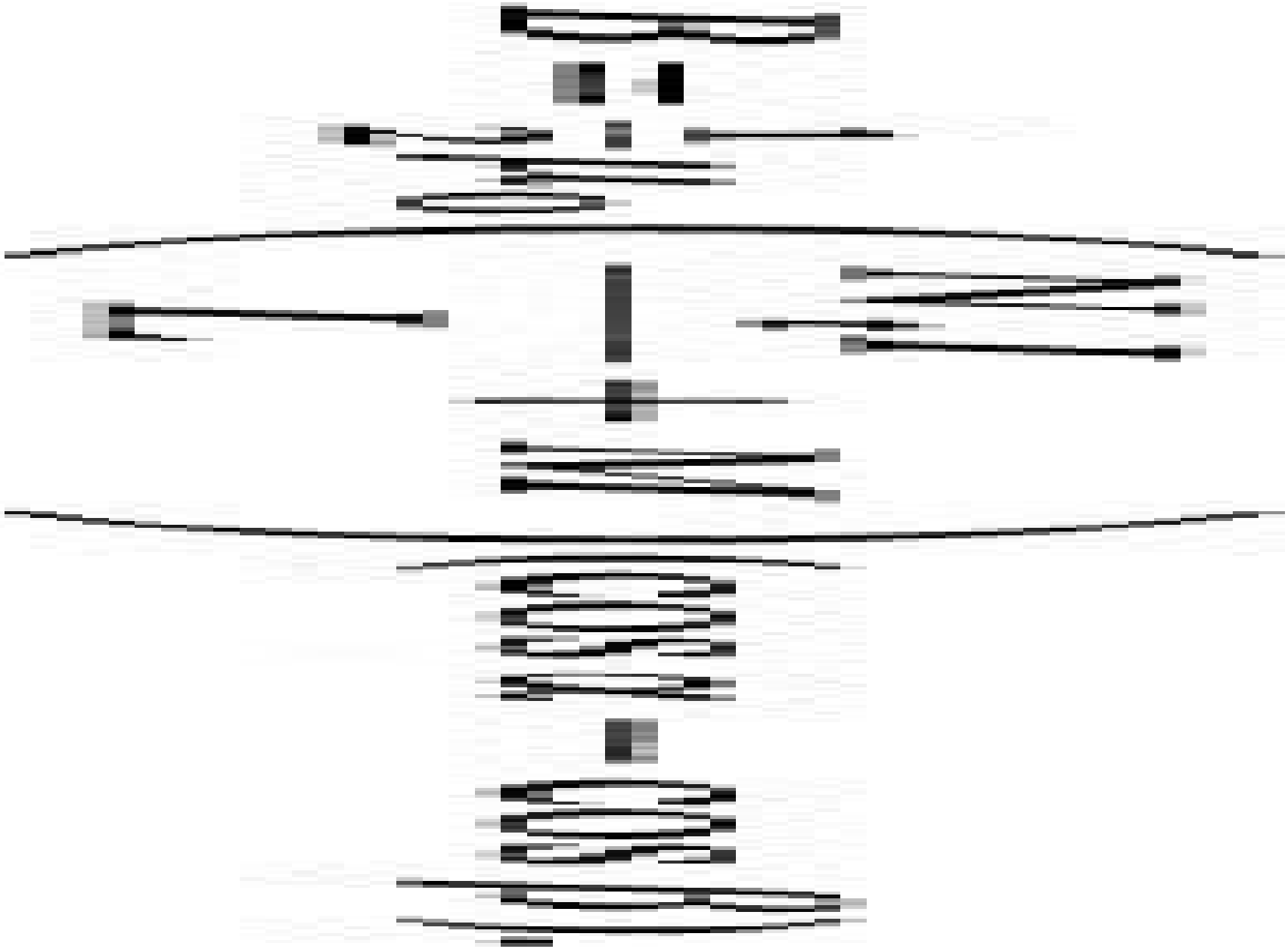
using the same symbols as in the text.



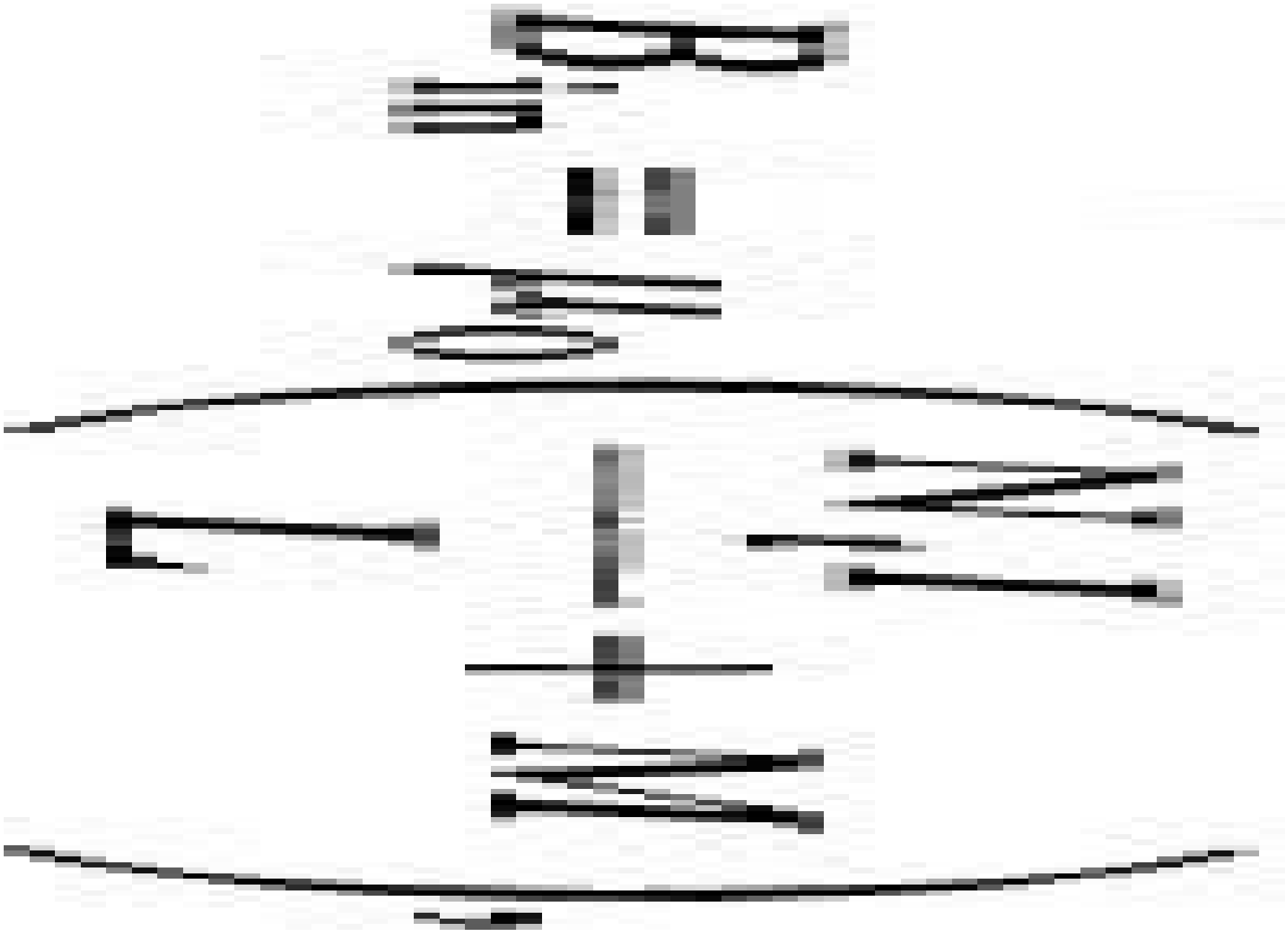
from the Biot–Savart law. The uniform magnetization of the rod is equivalent to a surface current IM = M per unit length around the surface of the rod. Hence



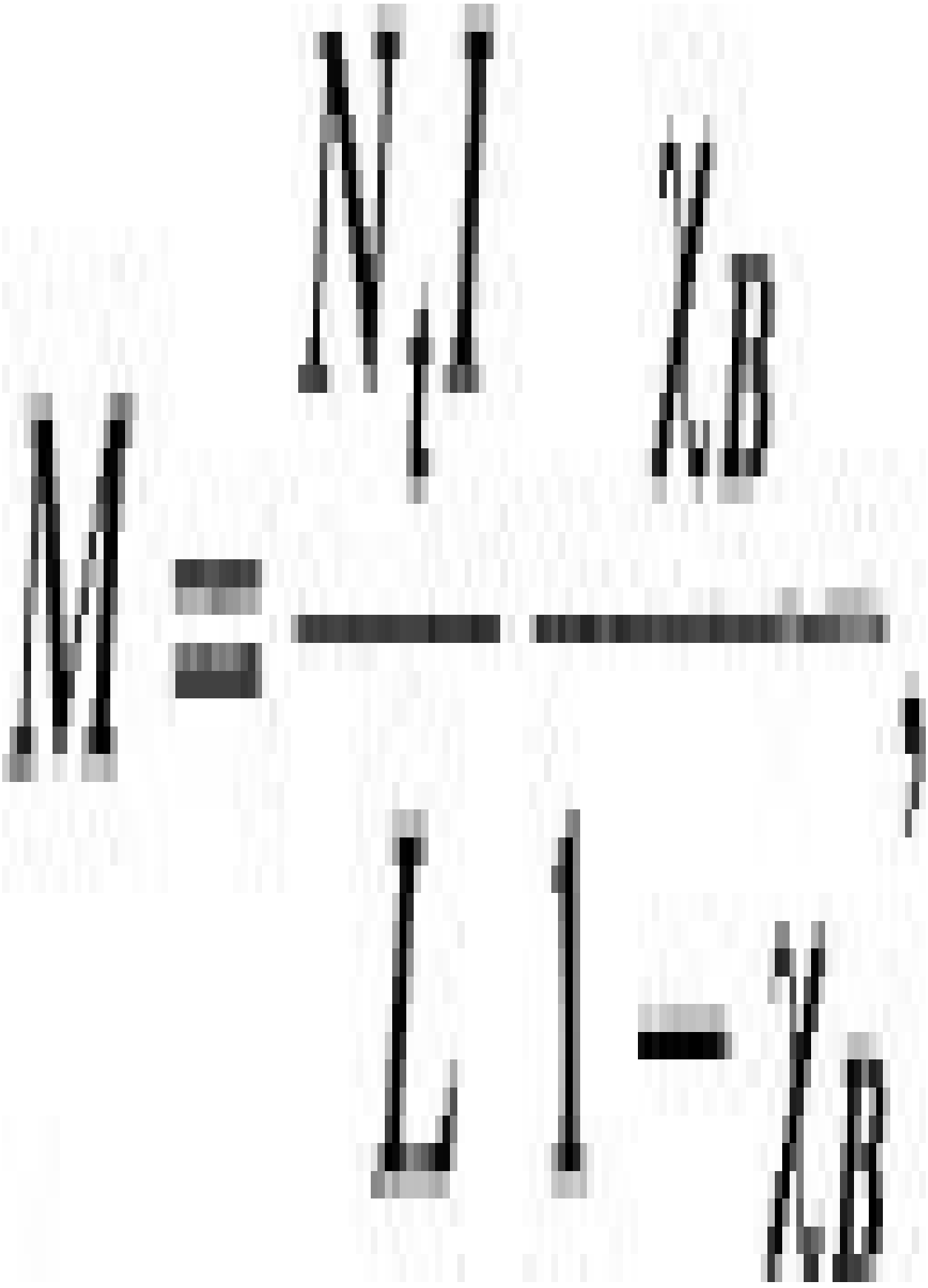
in the same way. The total field is therefore



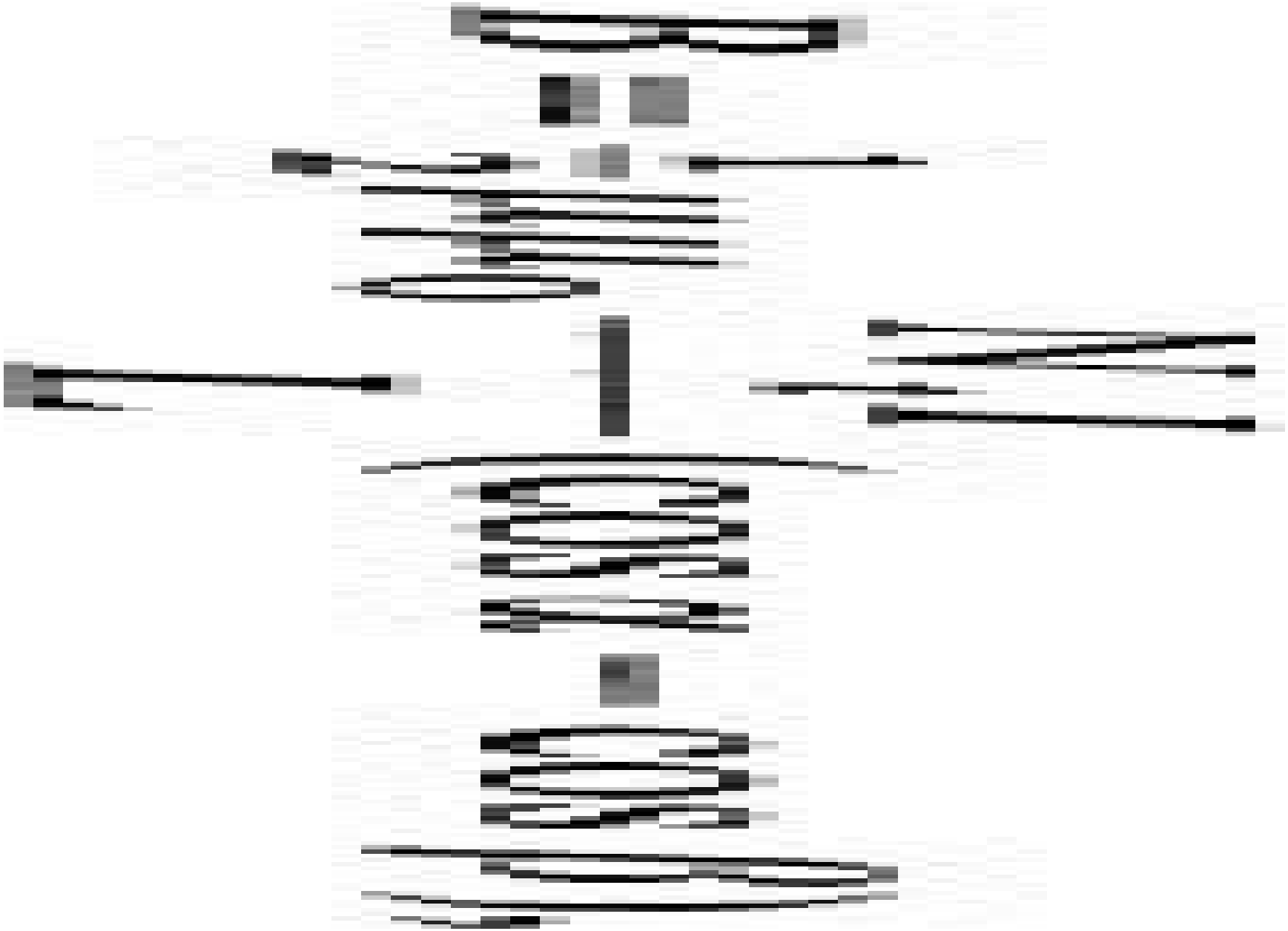
The field inside the solenoid to a good approximation is given by



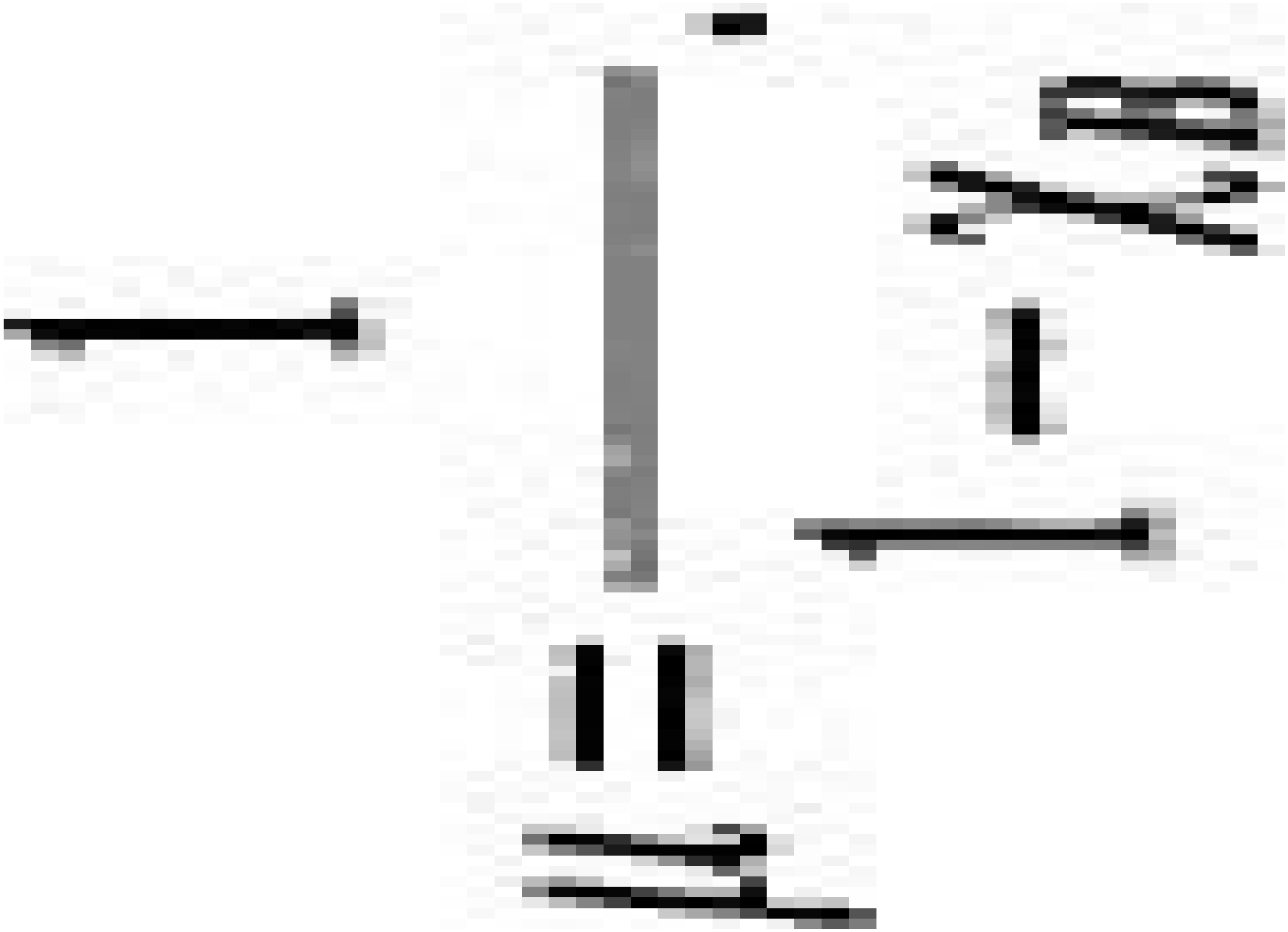
hence the uniform magnetization of the rod is



since $M = \chi\beta B_{in}/\mu_0$. The field B at the point P is now given by substituting for M in the above equation for B,



where



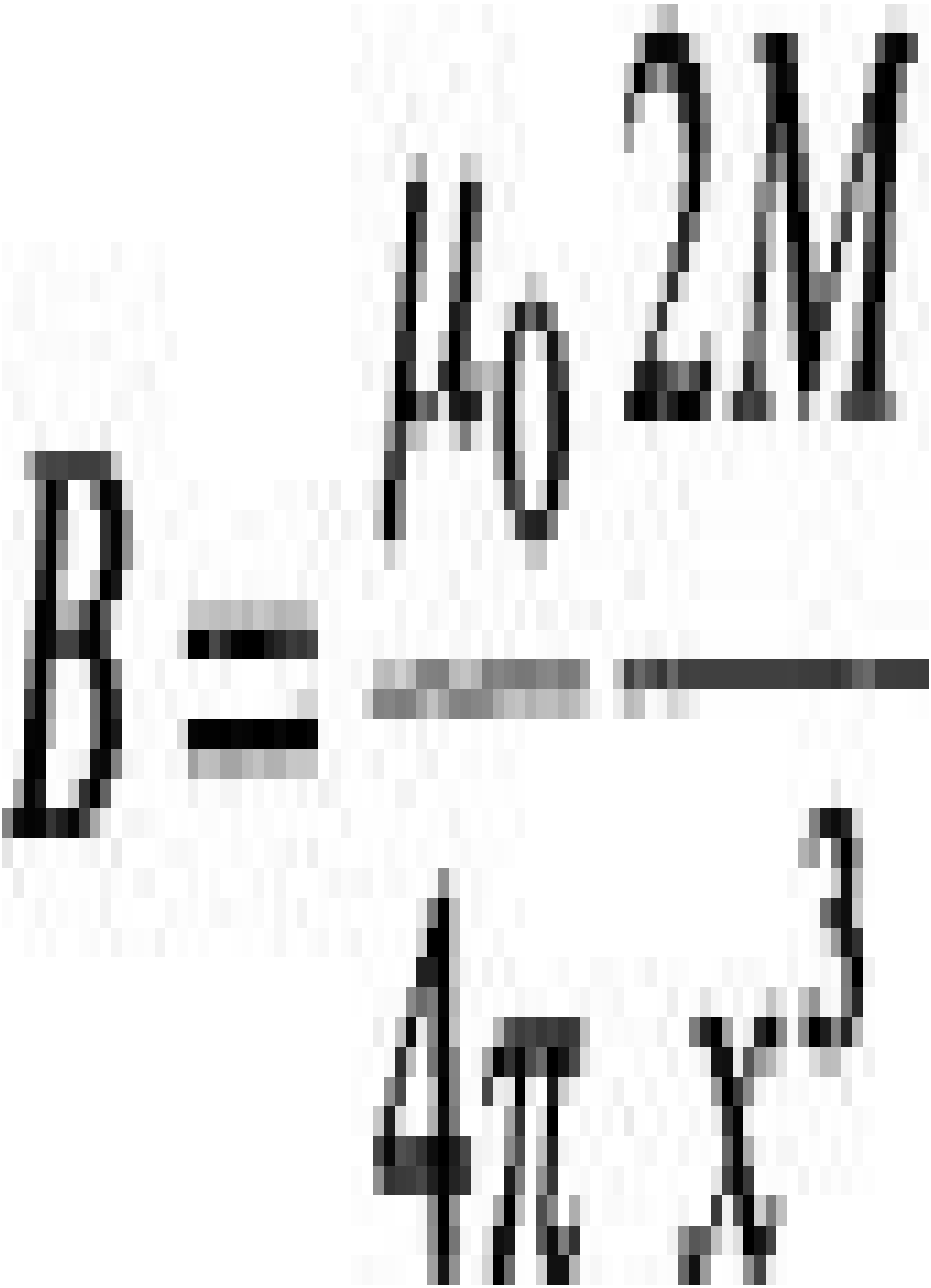
5.10 1.262 T.

5.11 2.48.

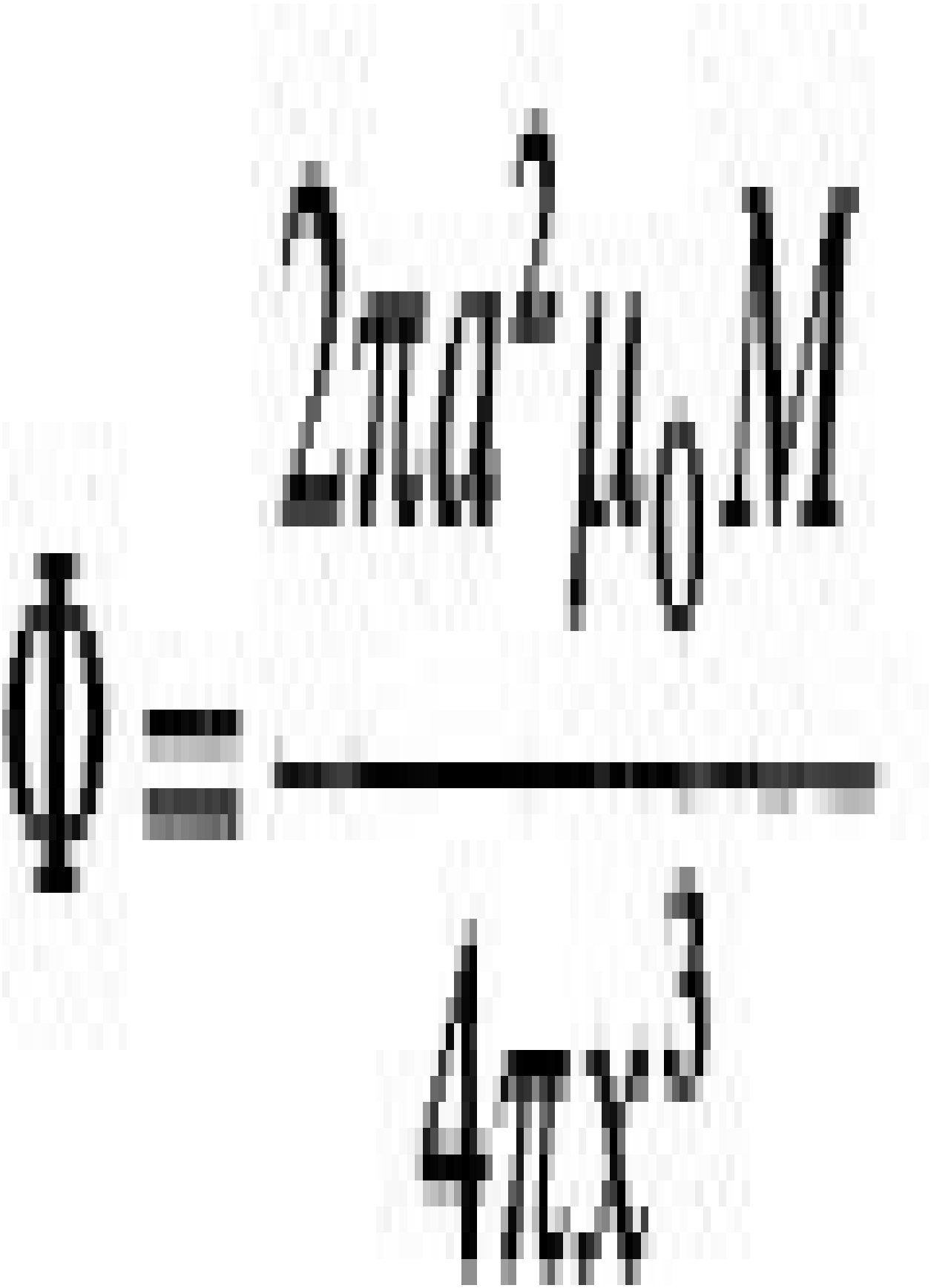
PROBLEMS 6

6.1 The force on a charge q a distance r from the axis of the rod is equal to qωrB in the direction of r increasing. This is equivalent to an electric field in this direction of magnitude ωrB. The polarization of the rod at distance r is thus P = χEε0ωBr, and this is equivalent to a volume distribution of charge −div P = −∂Pr/∂r − Pr/r = −2χEε0ωB throughout the rod, plus a surface distribution of charge σp = χEε0ωaβ over the surface of the rod. See Equations (2.3) and (2.17).

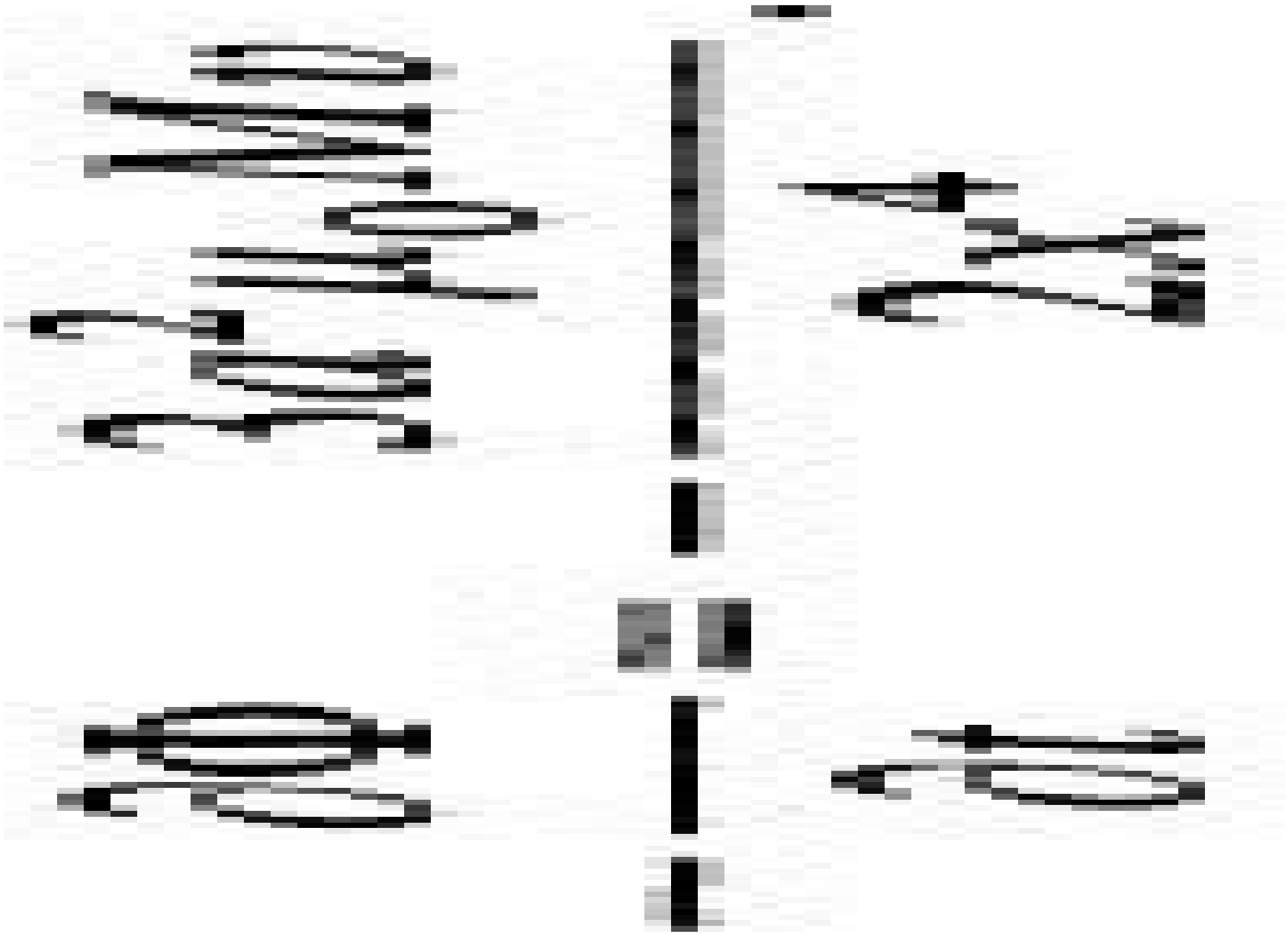
6.2 The field on axis due to the dipole M is given by Equation (4.29)



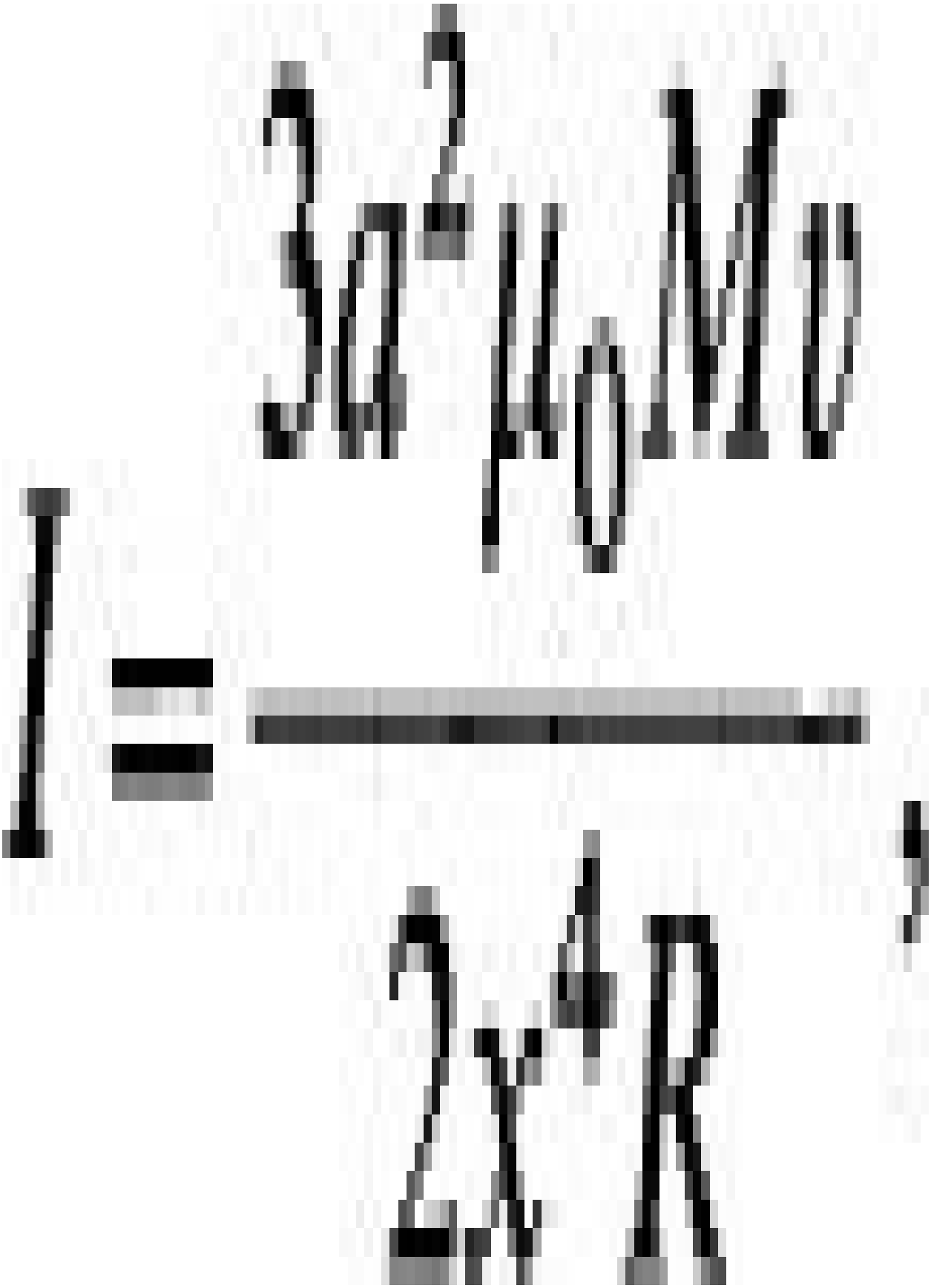
where x is the distance of the particle from the loop. The flux Φ through the loop is thus



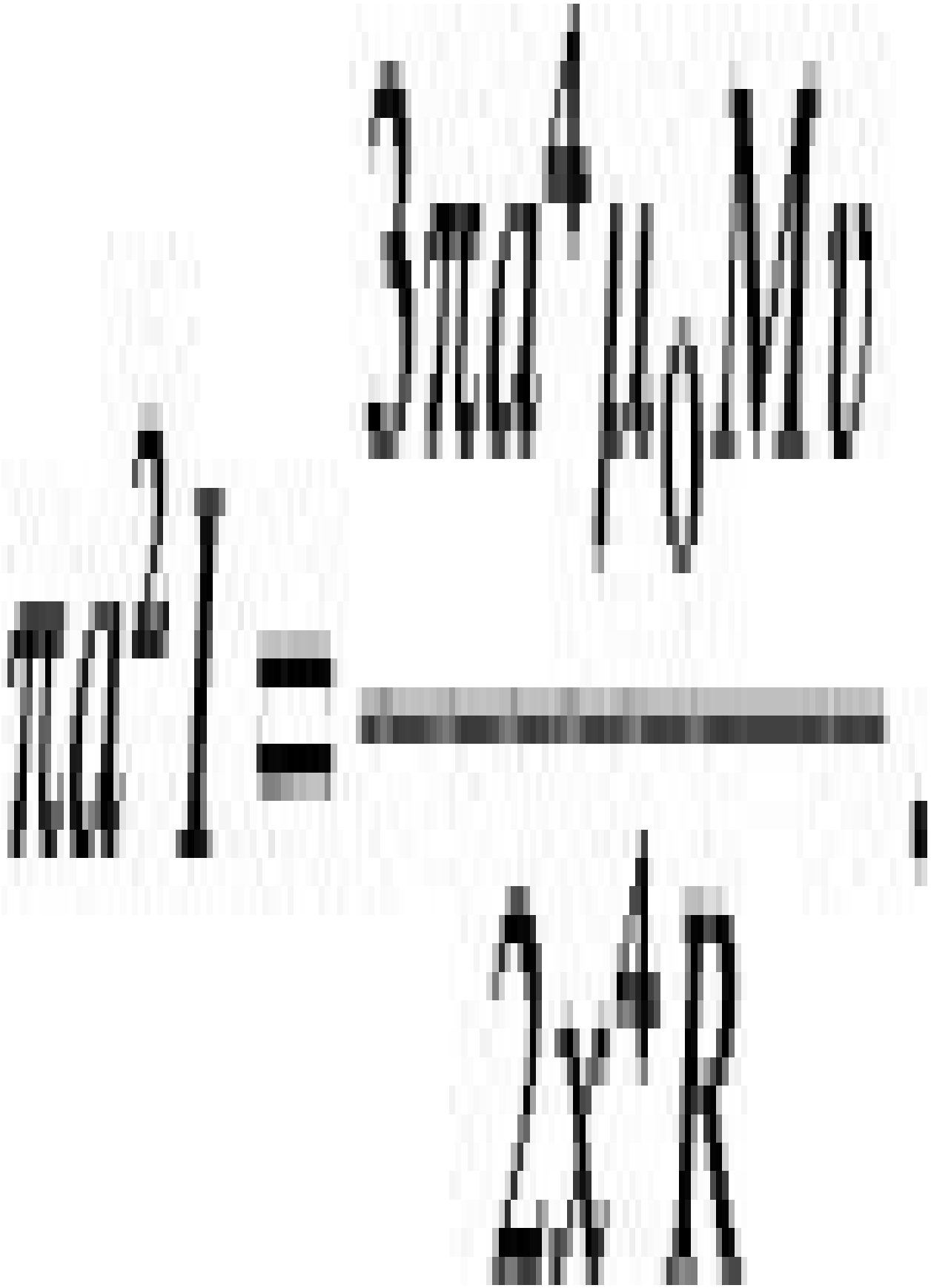
and the e.m.f. induced in the loop is



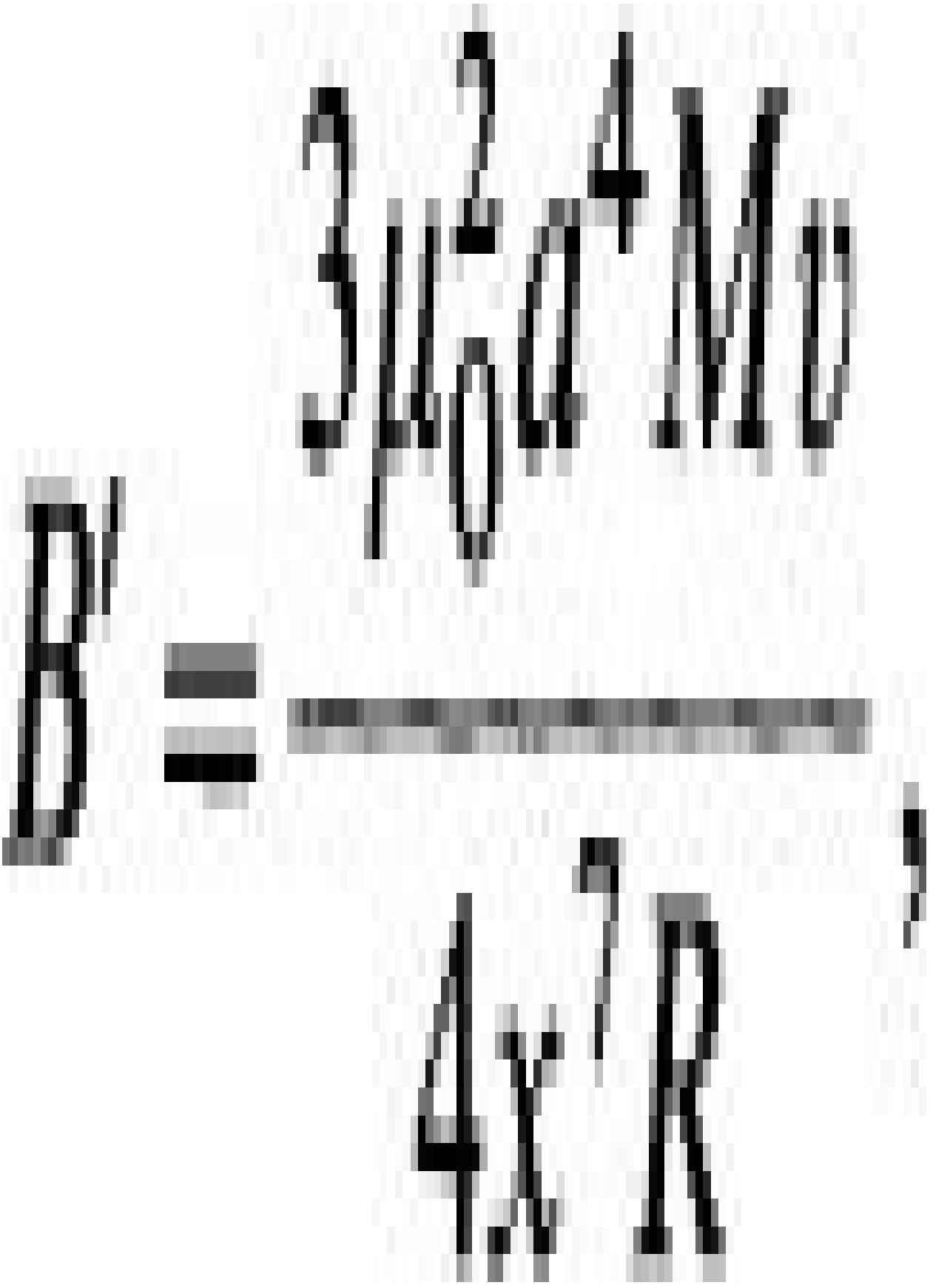
The current I in the loop is then



and the magnetic dipole moment associated with the current in the loop has the value

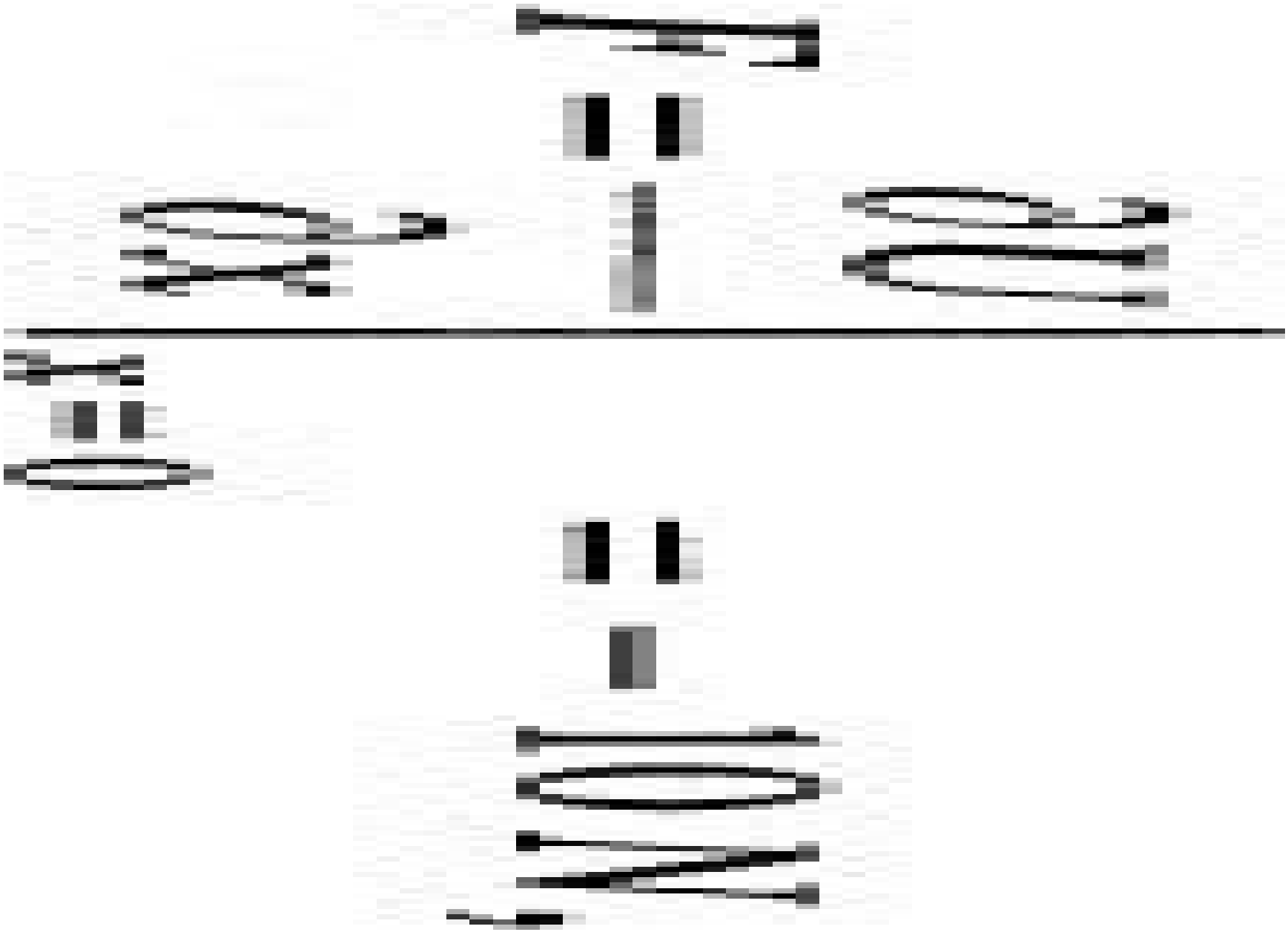


This produces a field at the position of the particle

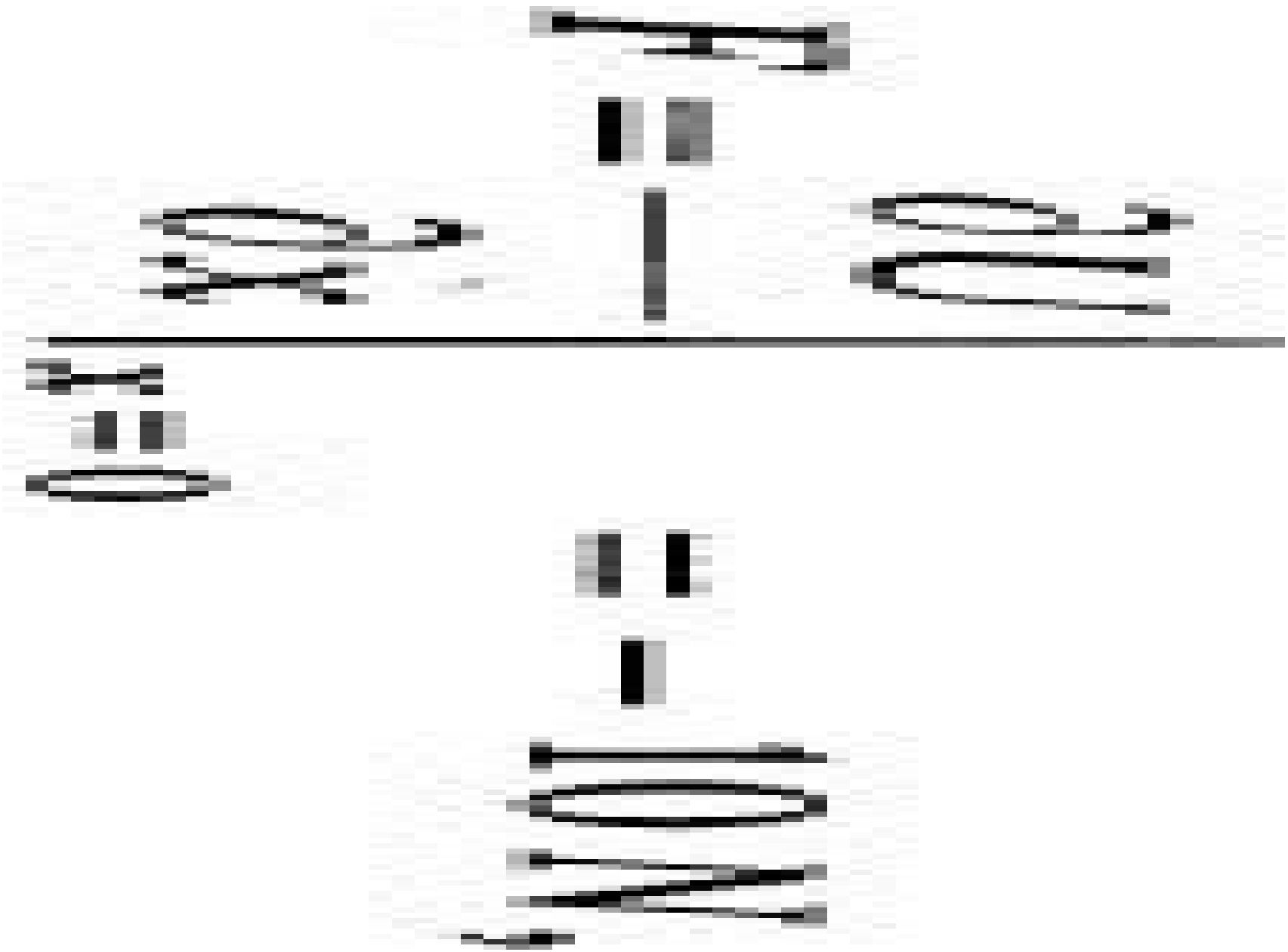


and the force on the particle is M∂B′/∂x, which is equal to 21 .

6.3 The force attracting the armature is equal to + grad U from Equation (6.33) where U = ∫vB.Hdτ is the total magnetic energy. This integral should be calculated for a separation x of the iron and current I through the coils. Then



and this gives the current I necessary to close the relay. The answer is 2.1 A. For the armature to open again, we put



giving a current of 0.08 A.

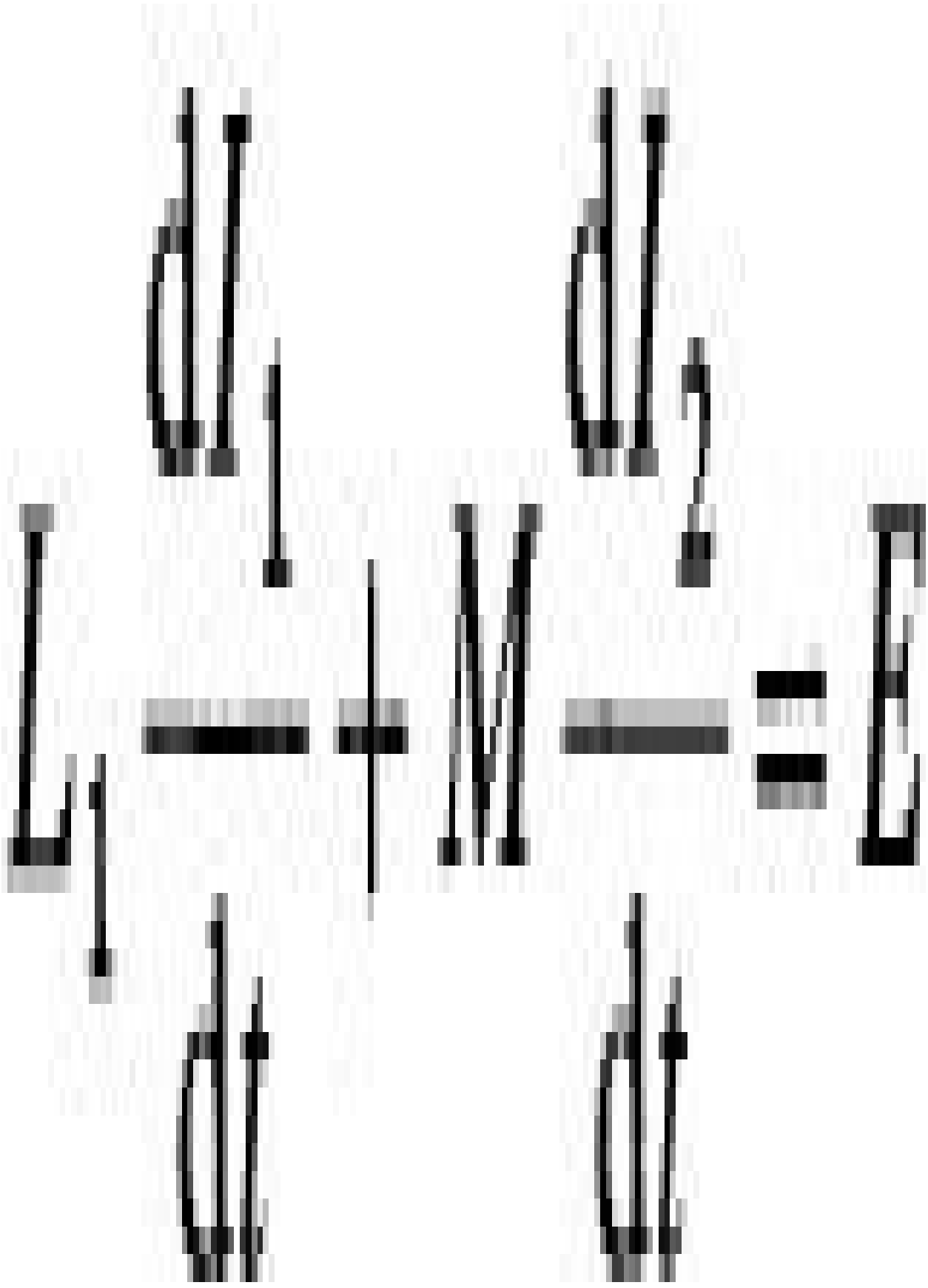
6.5 (i) Magnetic flux increased, hence self-inductance increases. There are some induced (eddy) currents flowing circularly around the axis of the bar. These tend to reduce the increase of flux and also dissipate energy, causing a slight increase in apparent resistance of the solenoid.

(ii) Large eddy currents flow in the high-conductivity copper, in such a direction that the flux change is reduced, leading to a smaller self-inductance. There is also a slight increase in apparent resistance.

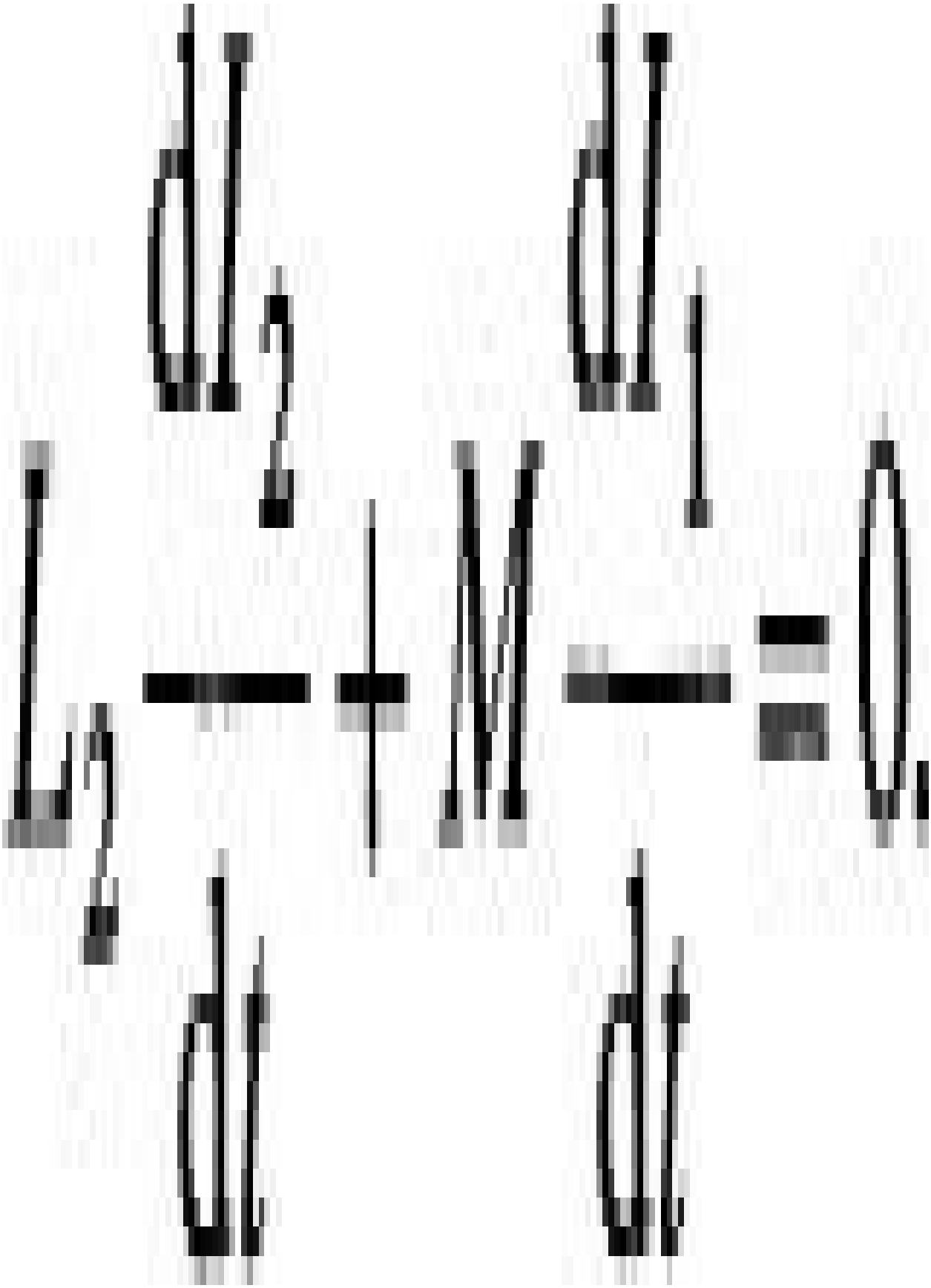
(iii) Similar to (ii).

(iv) No change in self-inductance or apparent resistance because induced currents cannot now flow.

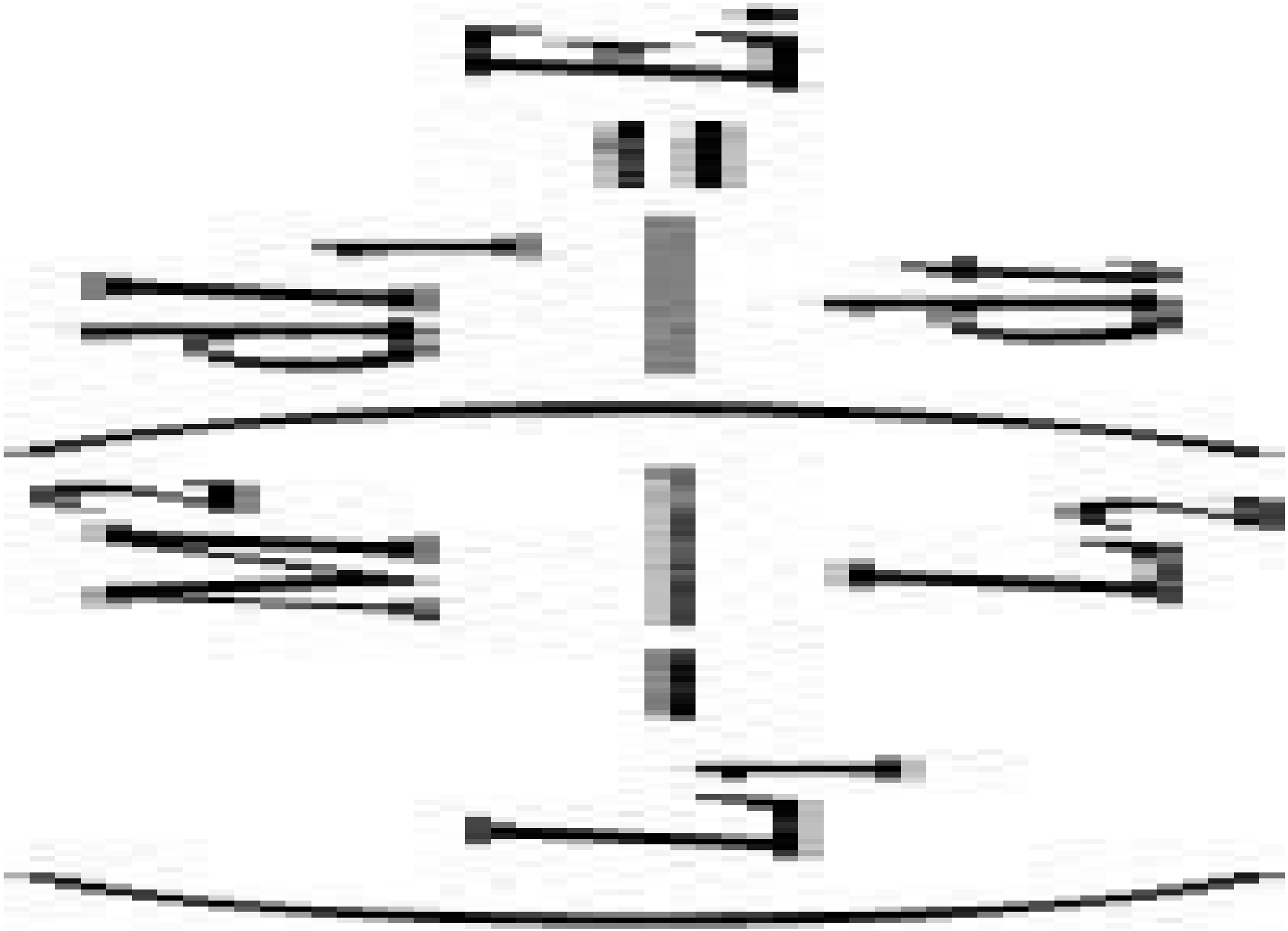
6.6 Initially the currents in both coils have to be zero, and the equations obtained by adding up voltages around the two coils at $t = 0$ are



and



These give the equation



6.7 The voltage difference between the rim and axis of the rotating flywheel shown in Figure D.2 is

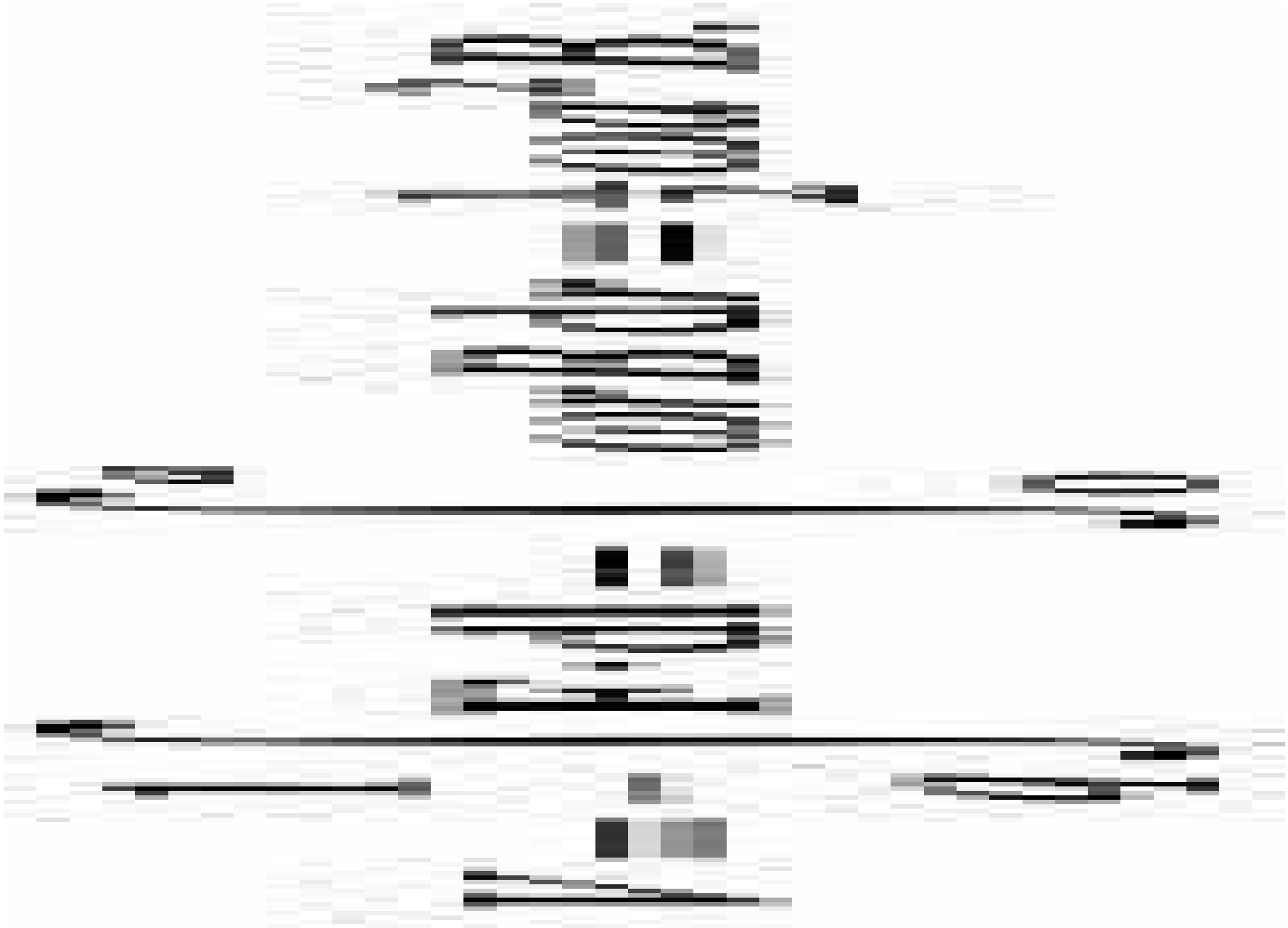
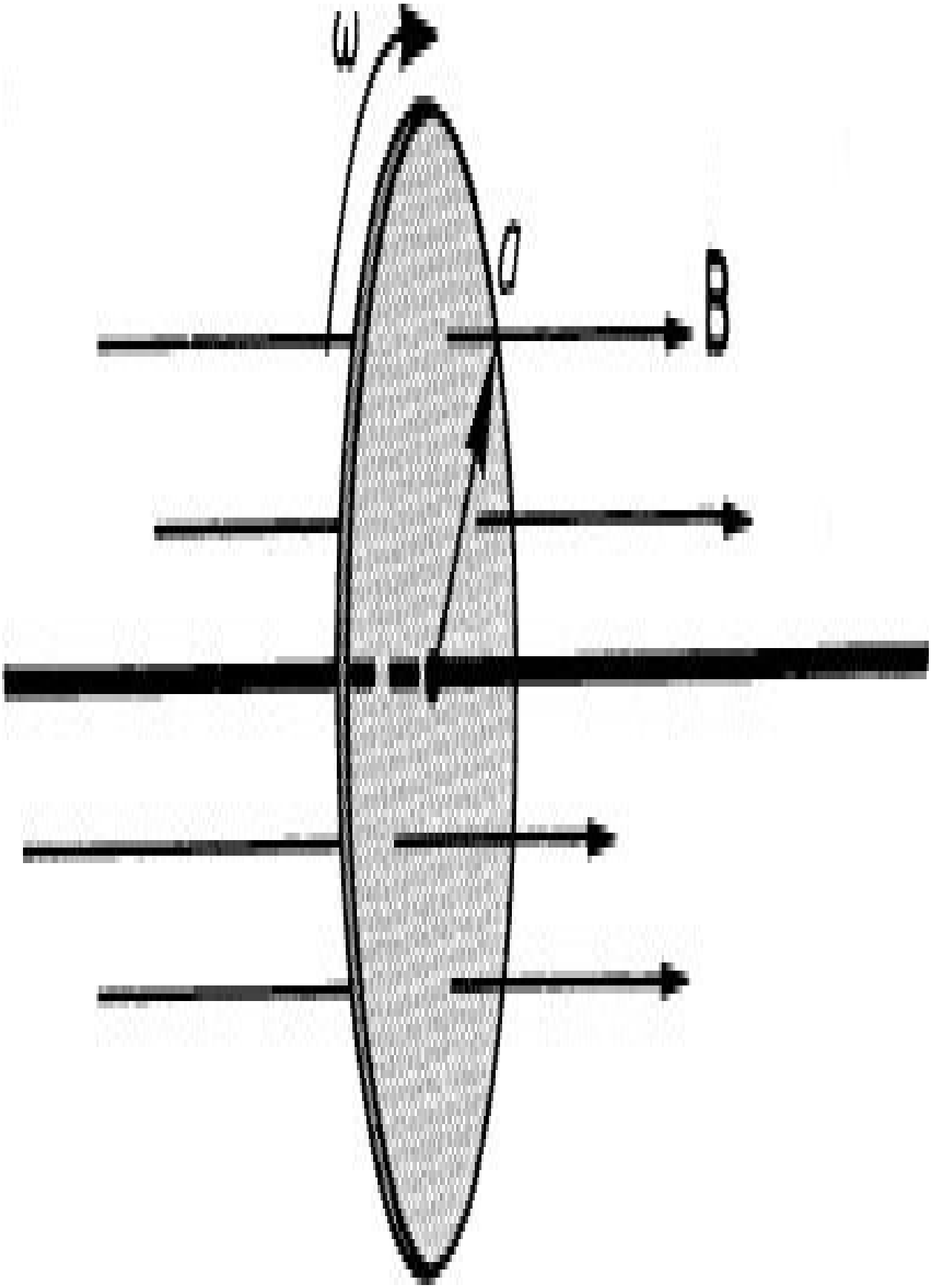
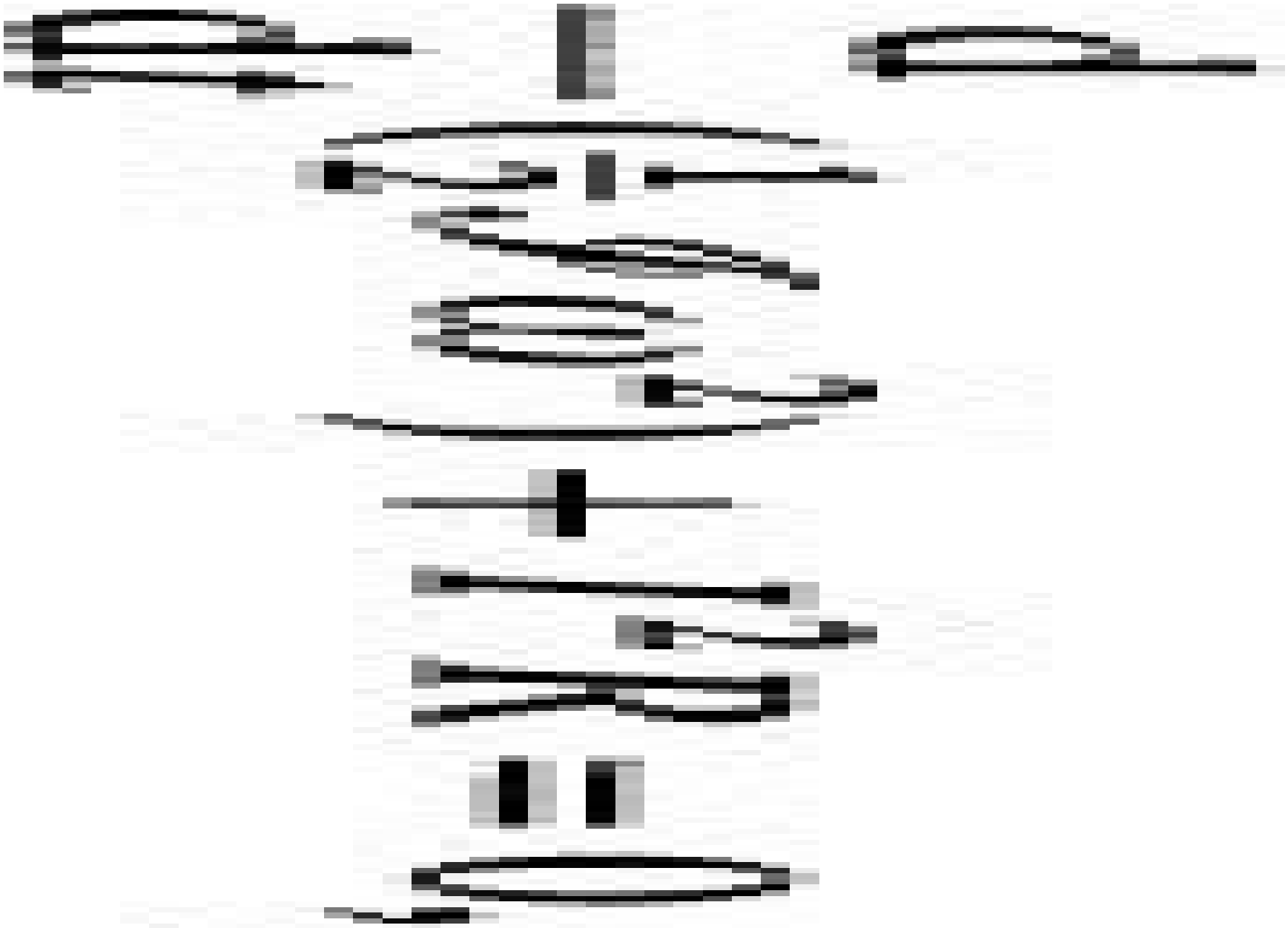


Figure D.2.

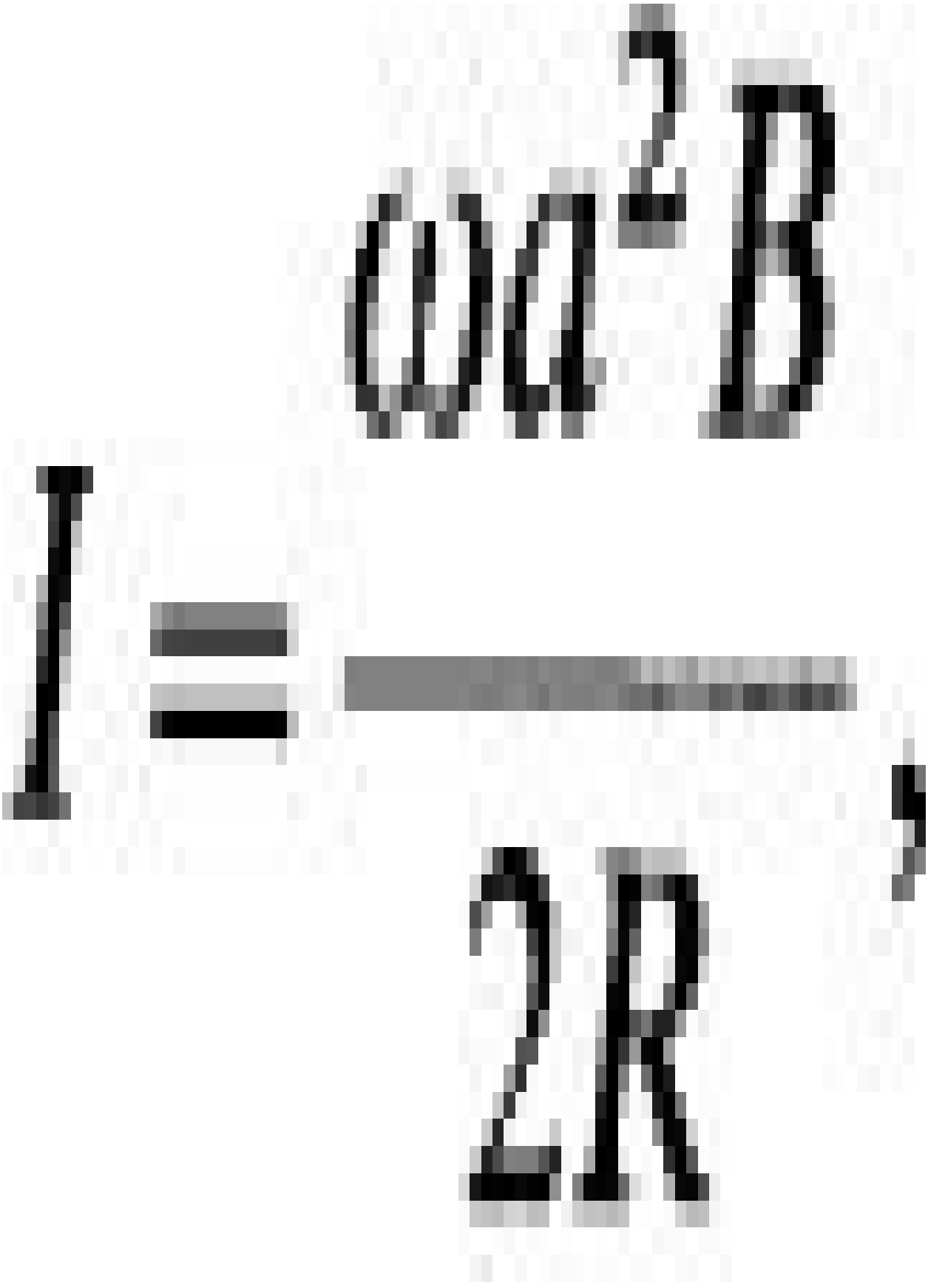

ω
α
B

When a load of 10−3Ω is connected the initial current is thus ωa²B/2R which in this example is $2.25\pi \times 10^5$ A.

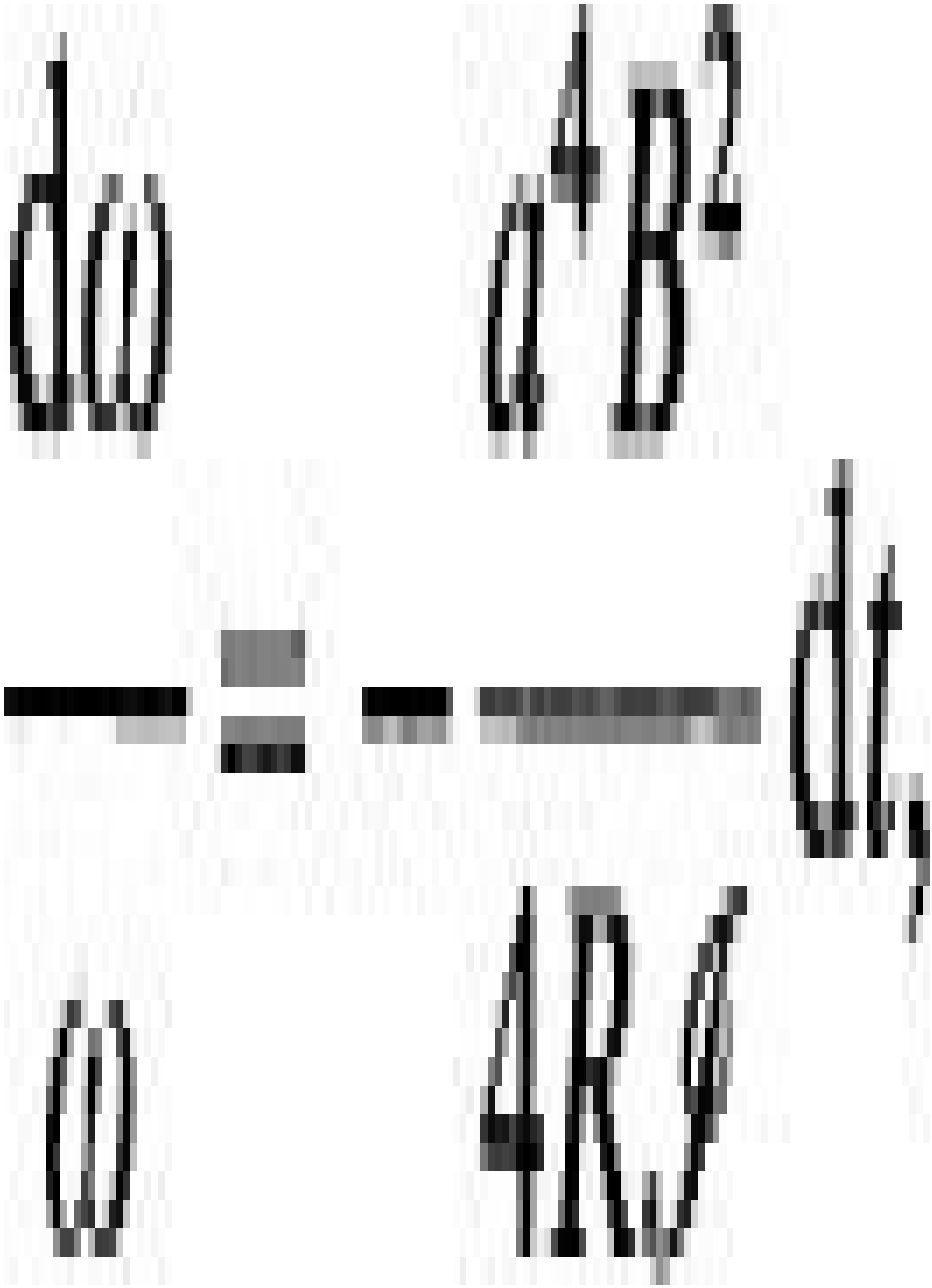
The rate at which the flywheel slows down can be calculated from the principle of conservation of energy which states that the rate of dissipation of heat in the load resistor is equal to the rate of loss of kinetic energy of the flywheel. Hence



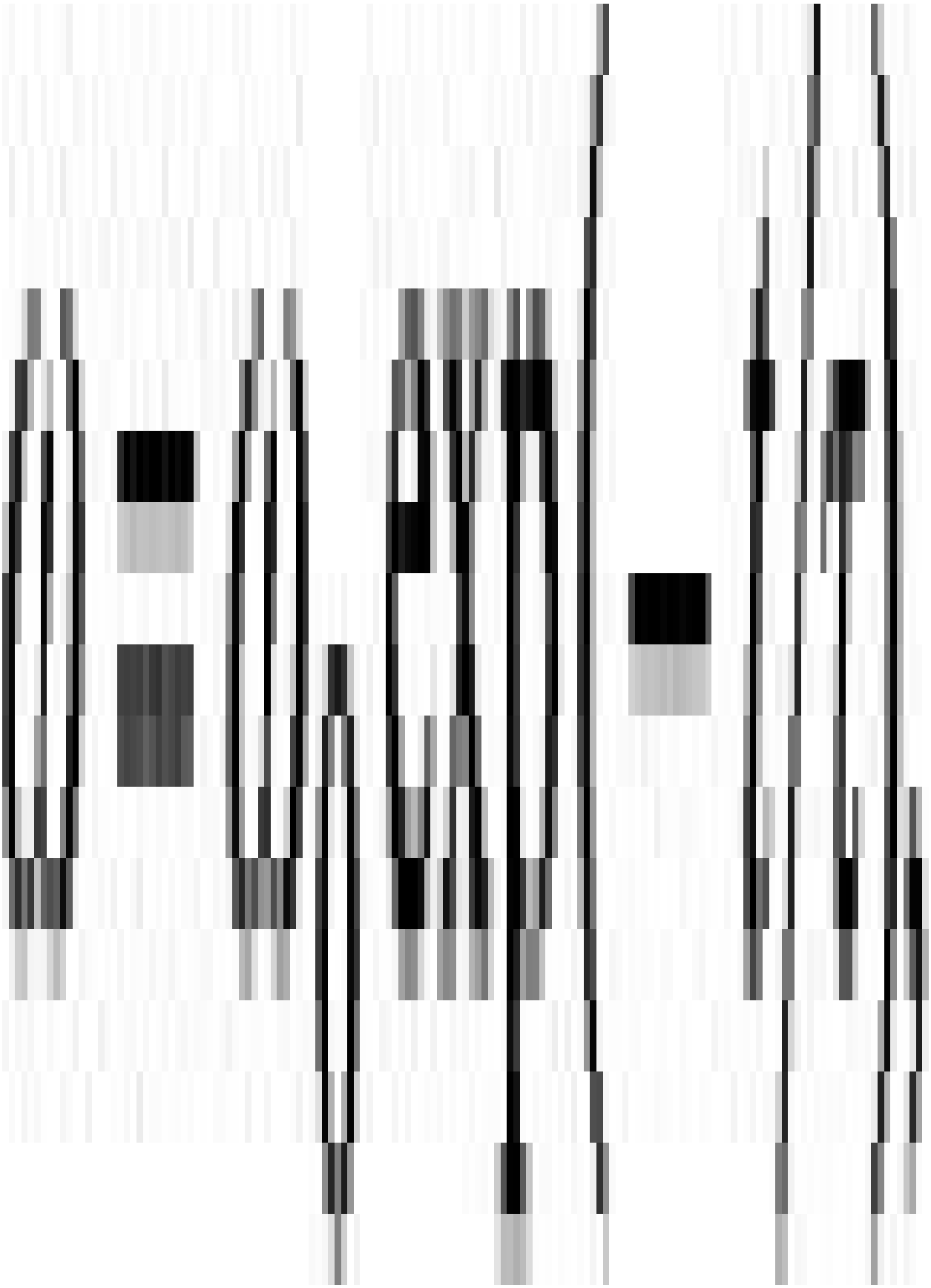
where is the moment of inertia of the flywheel, I the current and ω the angular frequency at time t.



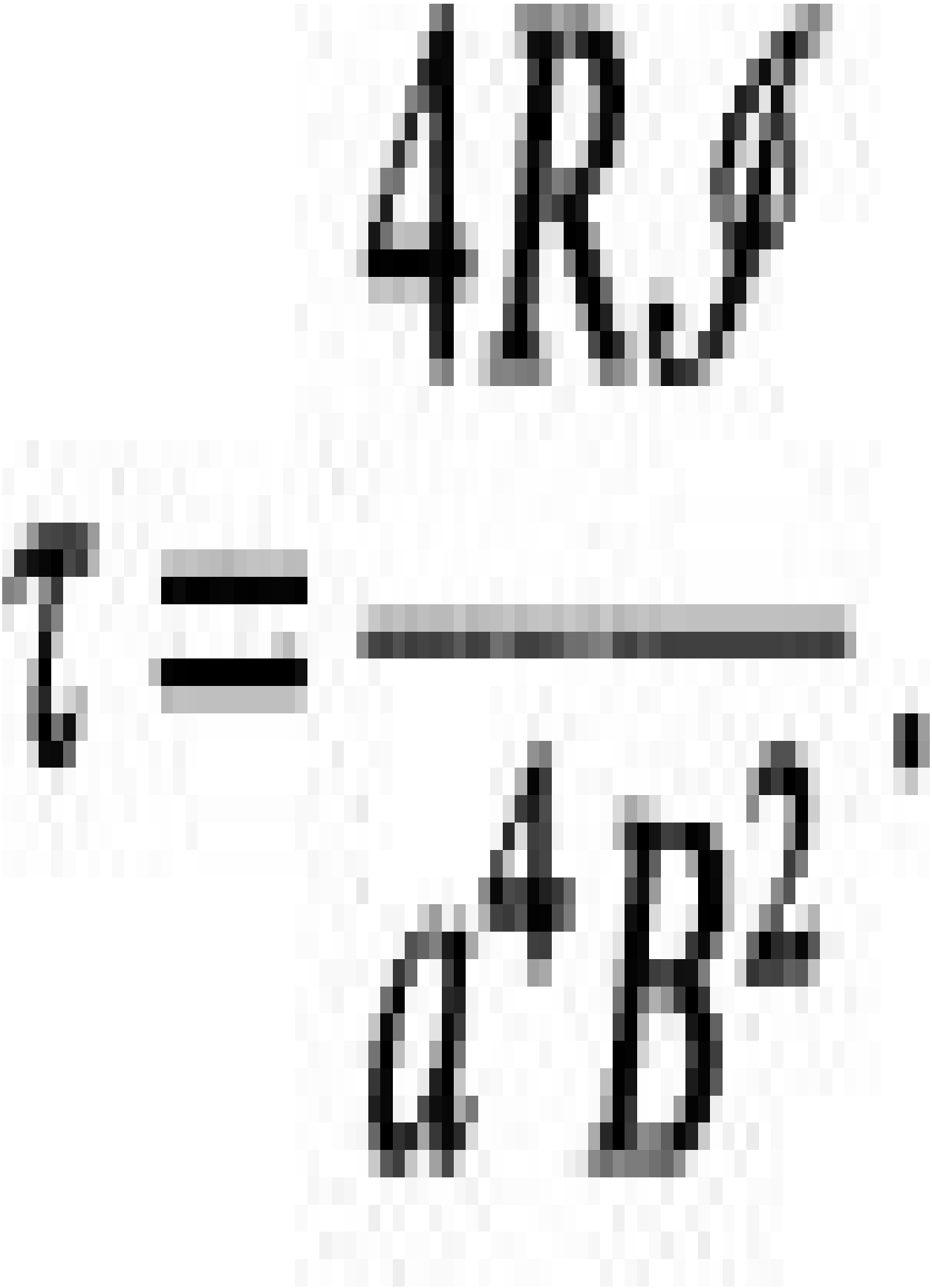
and so



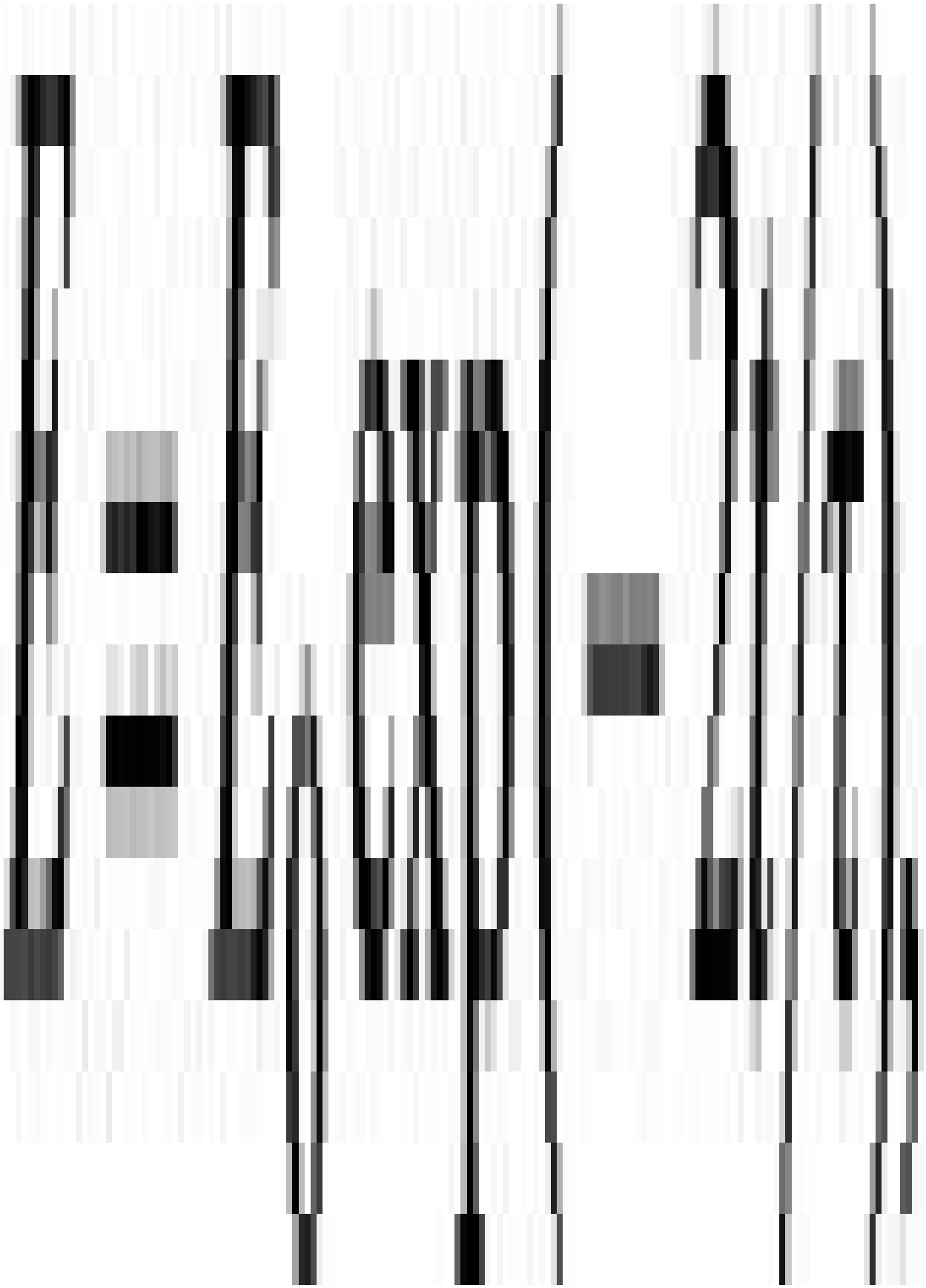
from which



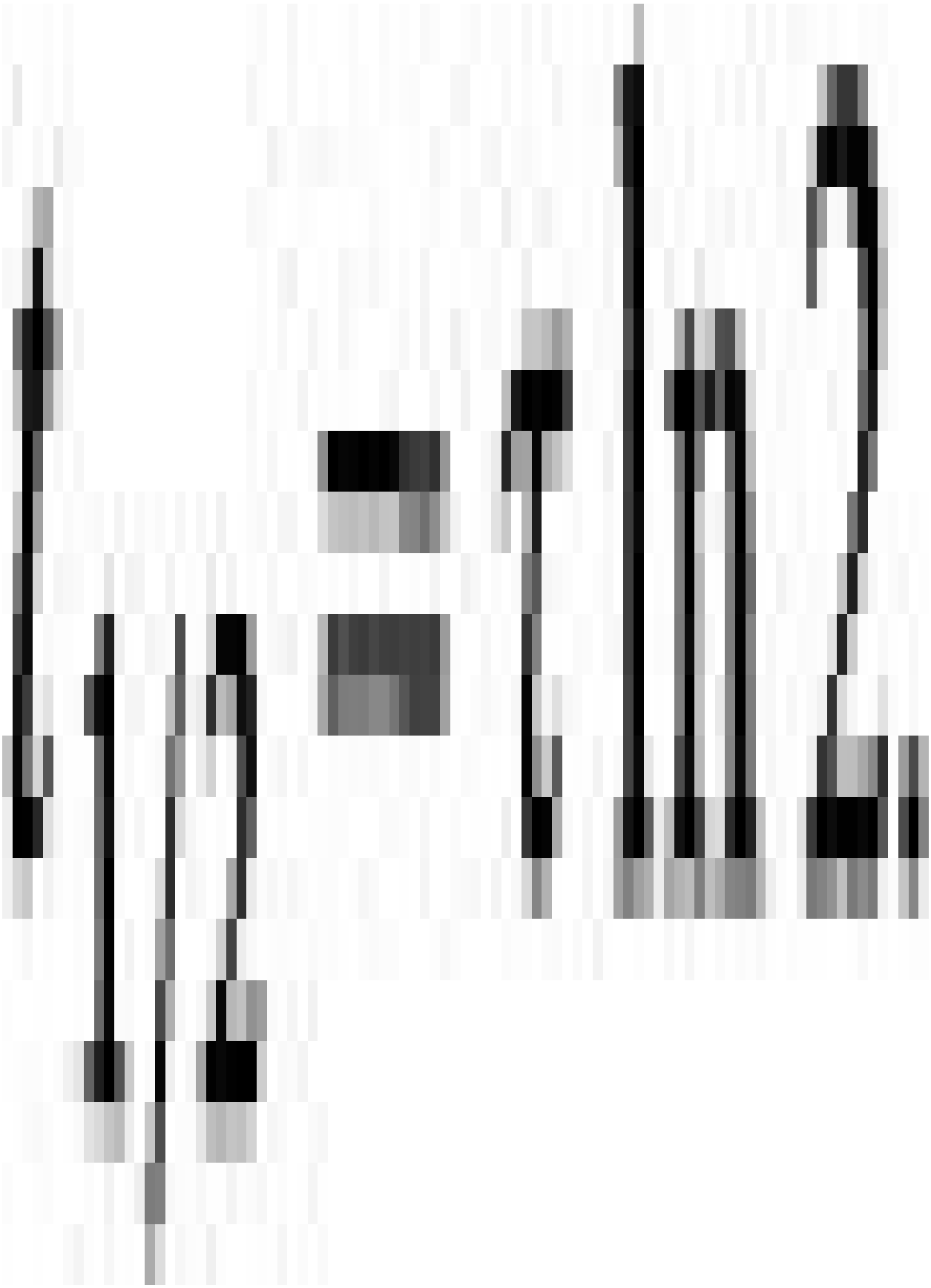
where



The energy of the flywheel at time t is thus given by



where E0 is the initial energy. The angular frequency thus falls to one half of its initial value in a time t1/2 where



In this example t1/2 is 6.16 seconds.

6.9 The flux Φ through the coil (Figure D.3) is

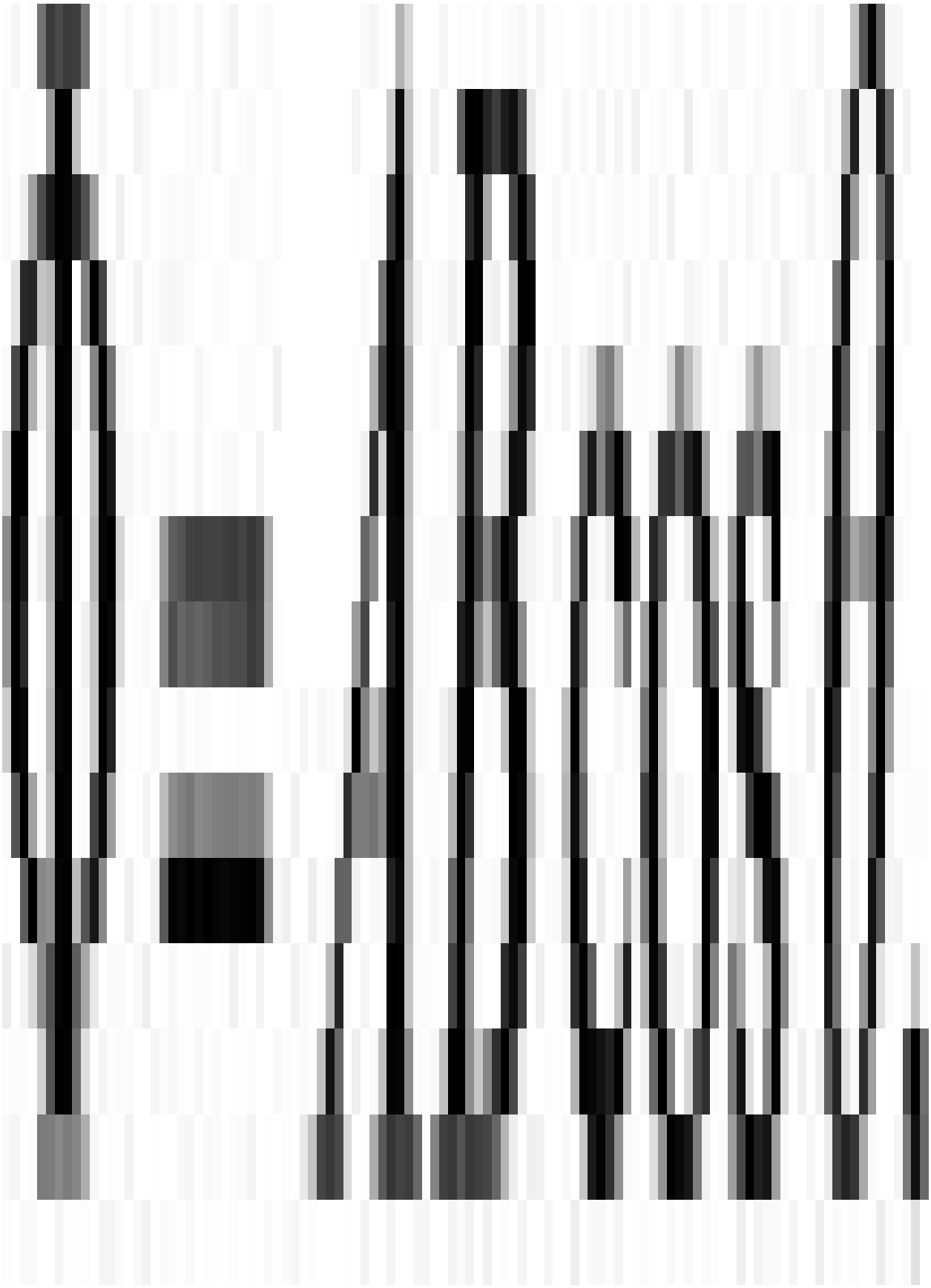
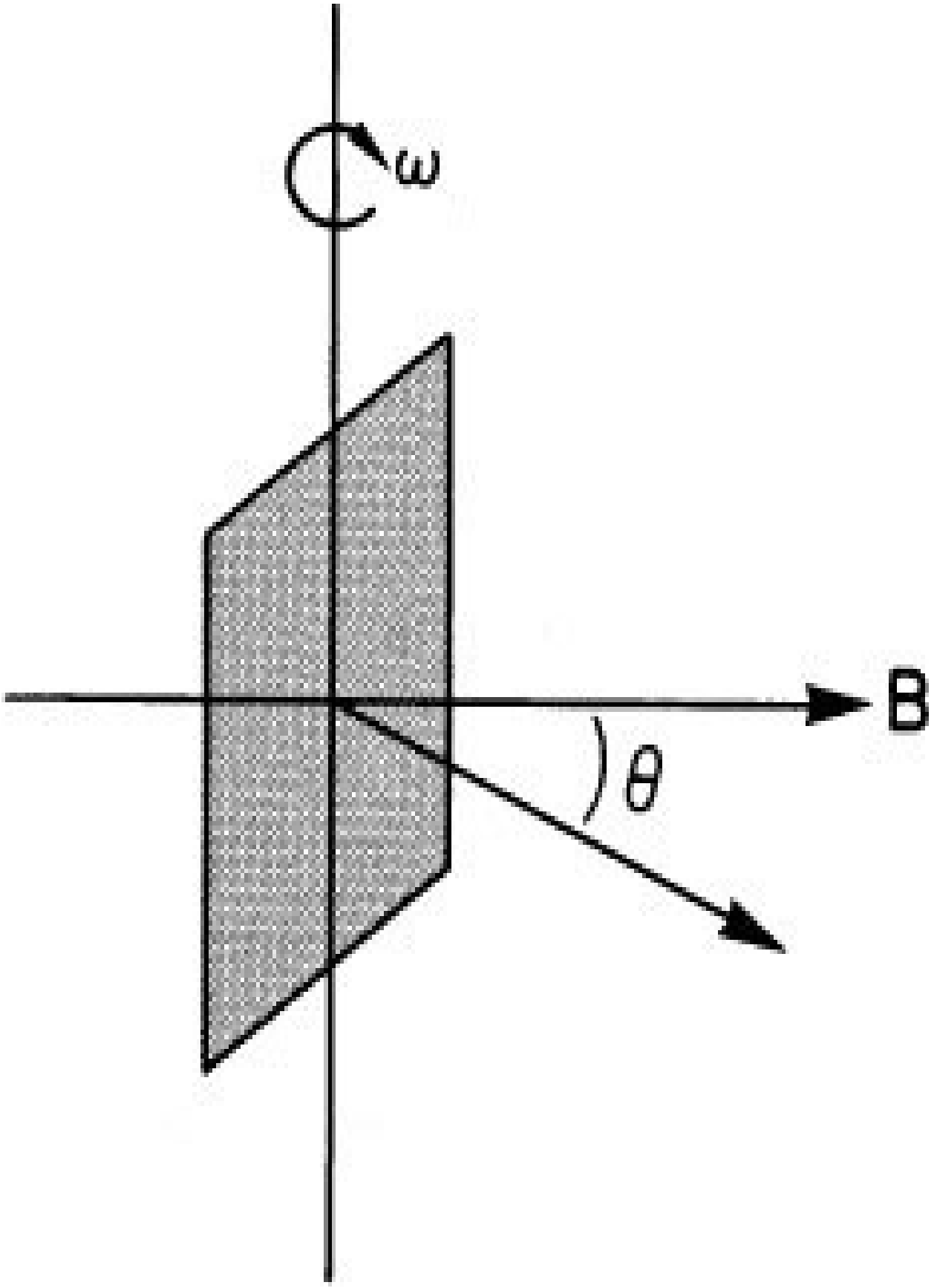
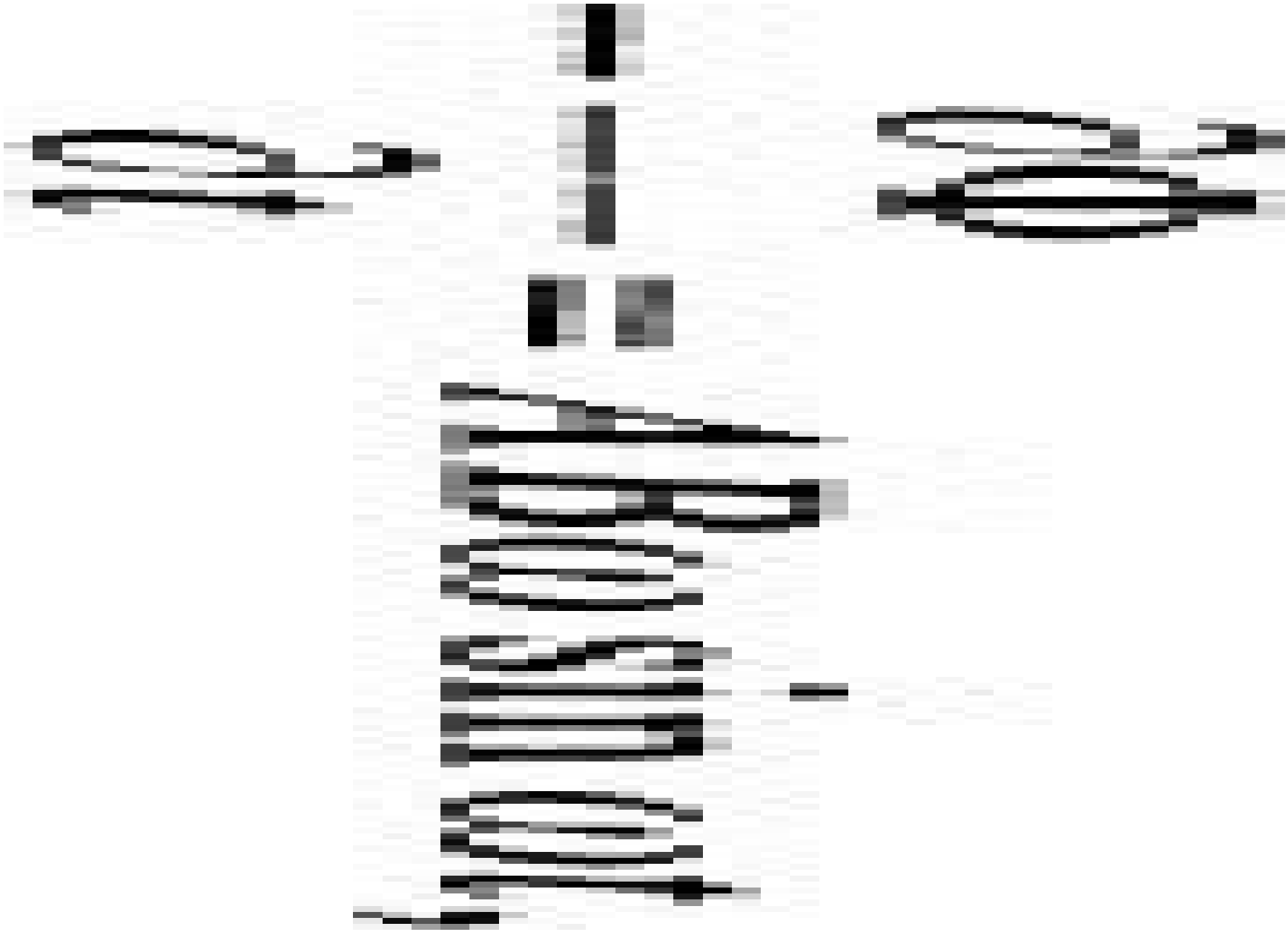


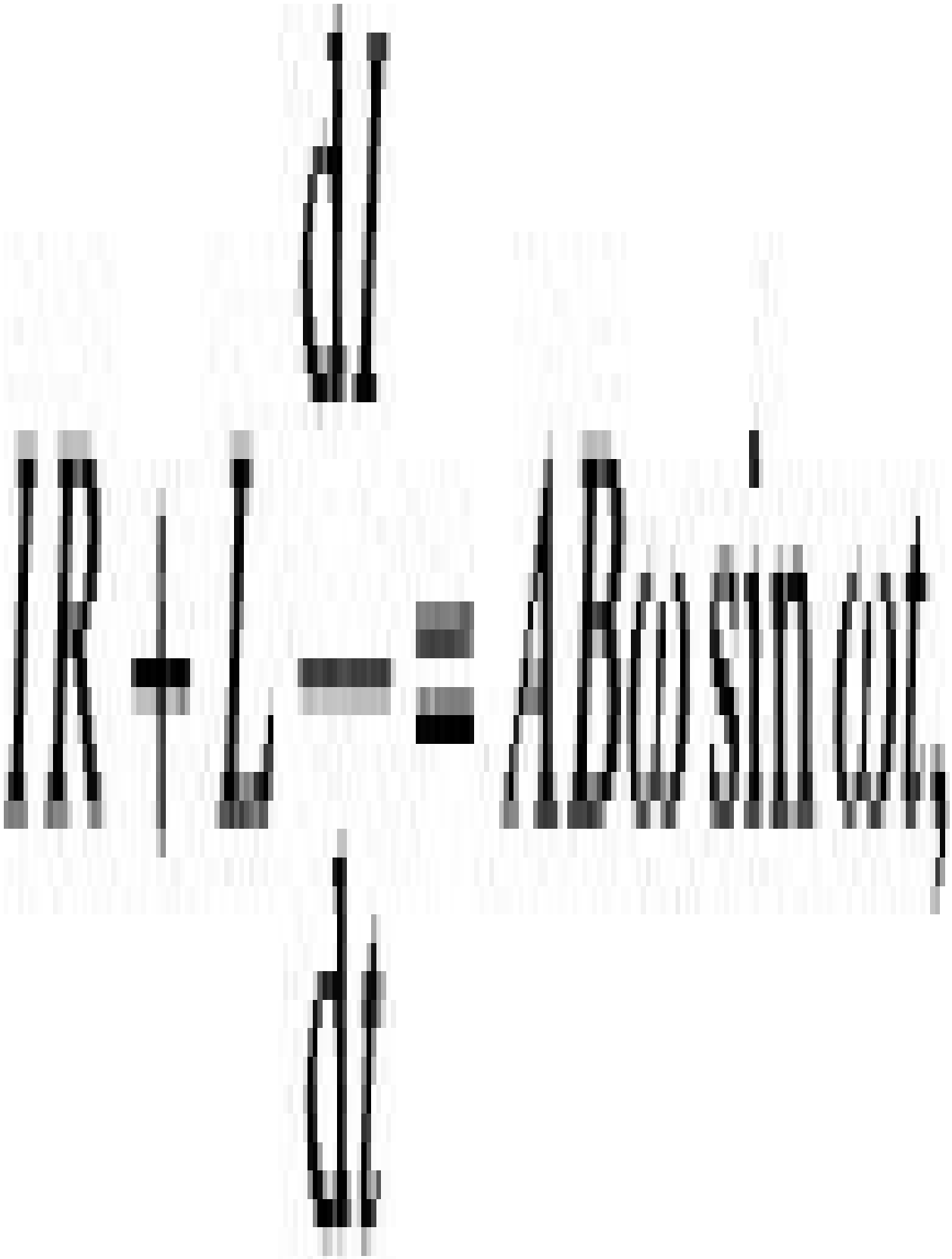
Figure D.3.


ω
B
θ

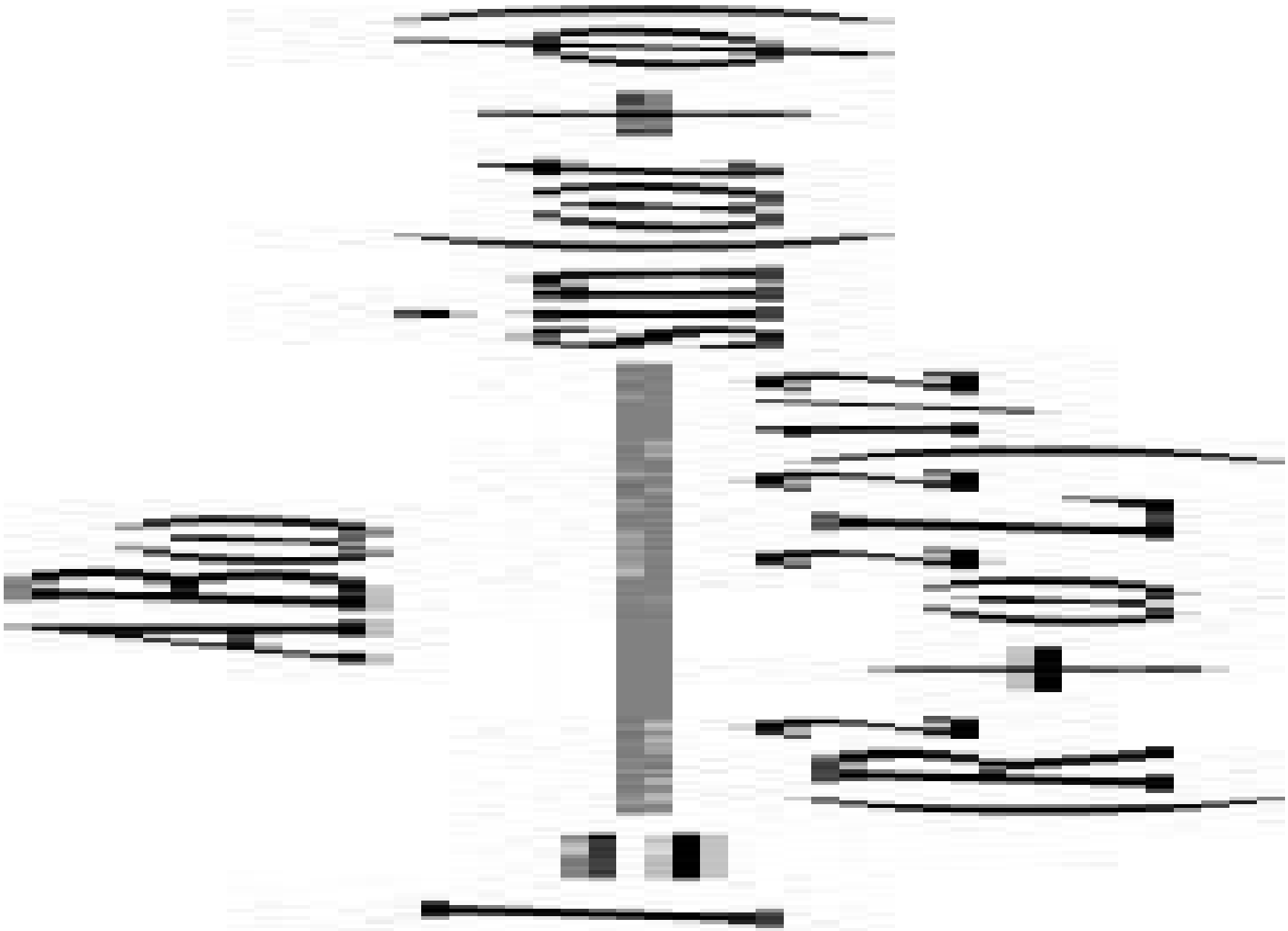
The e.m.f. induced in the coil as it rotates is



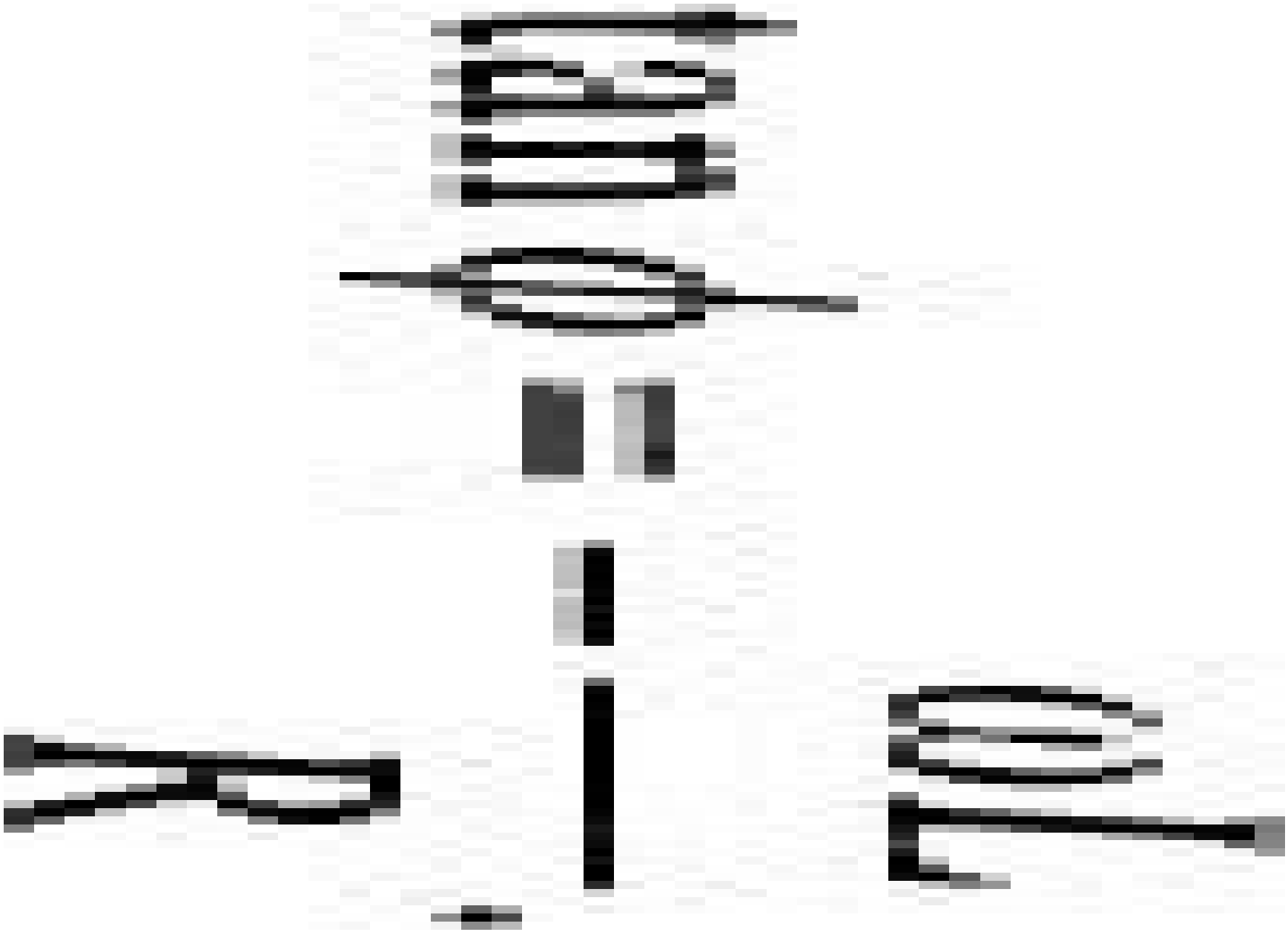
where ω is the angular speed of rotation. Adding the voltages around the coil we obtain



from which the current I is found to be



with



The energy dissipated in the resistance of the coil per cycle is $(AB\omega)^2 R/2(R^2 + \omega^2 L^2)$, and this is equal to $2\pi T$, where T is the mean torque required to rotate the coil.

6.11 χ_B is 1.2×10^{-5}.

6.13 Since

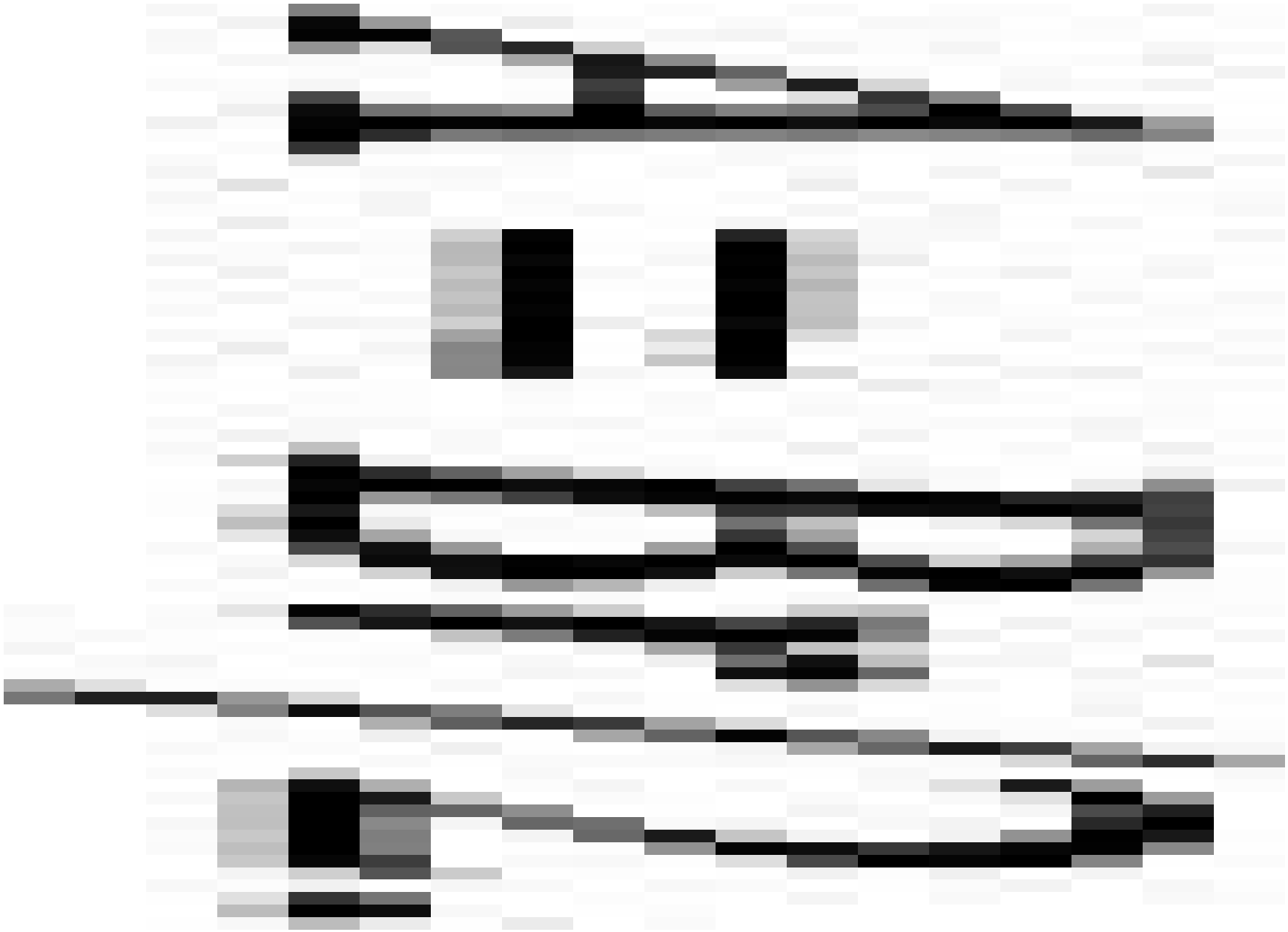
$$\mathbf{B} = \operatorname{curl} \mathbf{A},$$

$$\oint_s \mathbf{A} \cdot d\mathbf{l} = \int_S \operatorname{curl} \mathbf{A} \cdot d\mathbf{S},$$

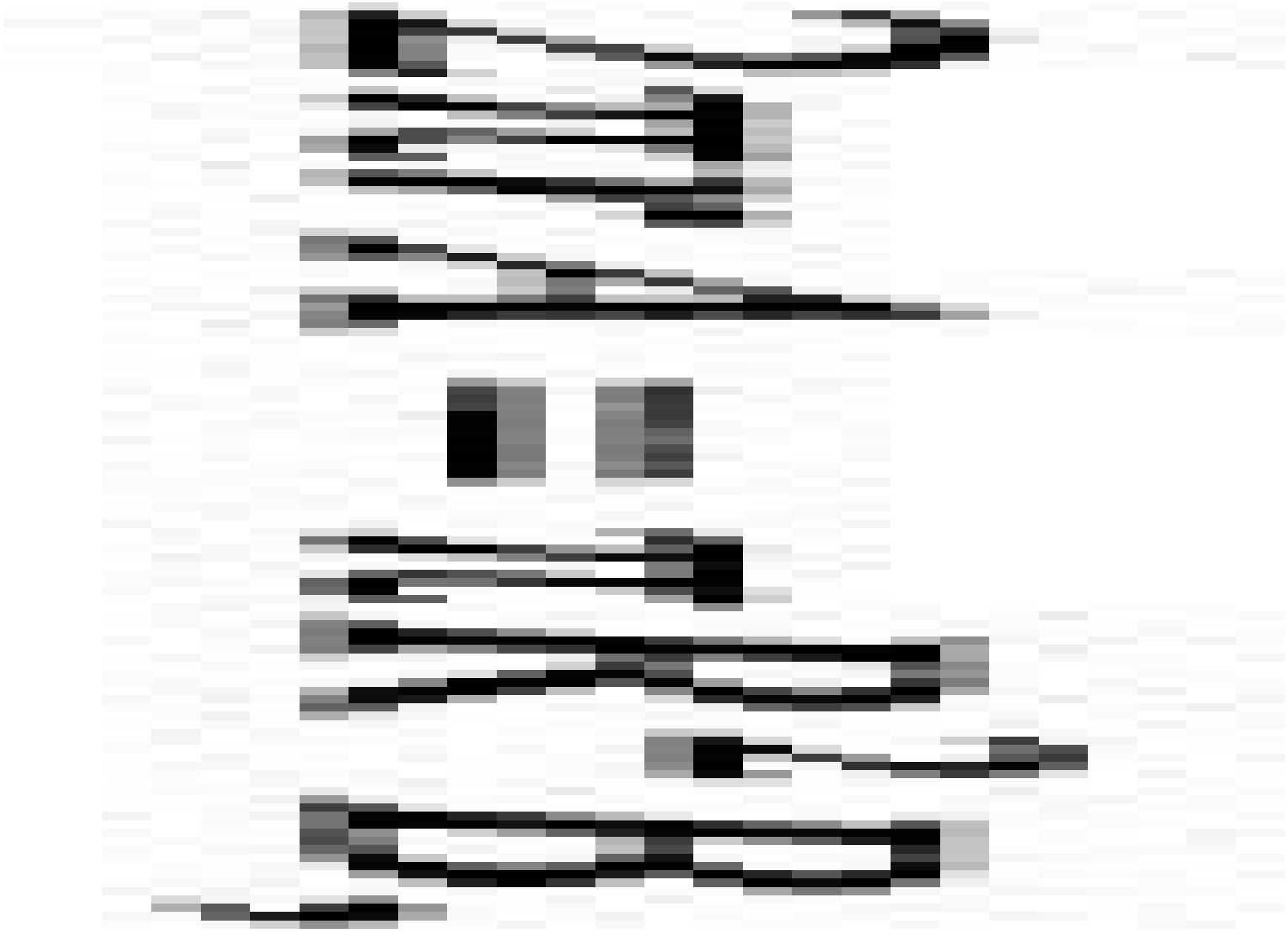
$$= \int_S \mathbf{B} \cdot d\mathbf{S}.$$

The symmetry of the situation requires that the potential A depend only on perpendicular distance from the solenoid axis r, and be tangential to coaxial circles of radius r. Around such a circle drawn inside the solenoid we have

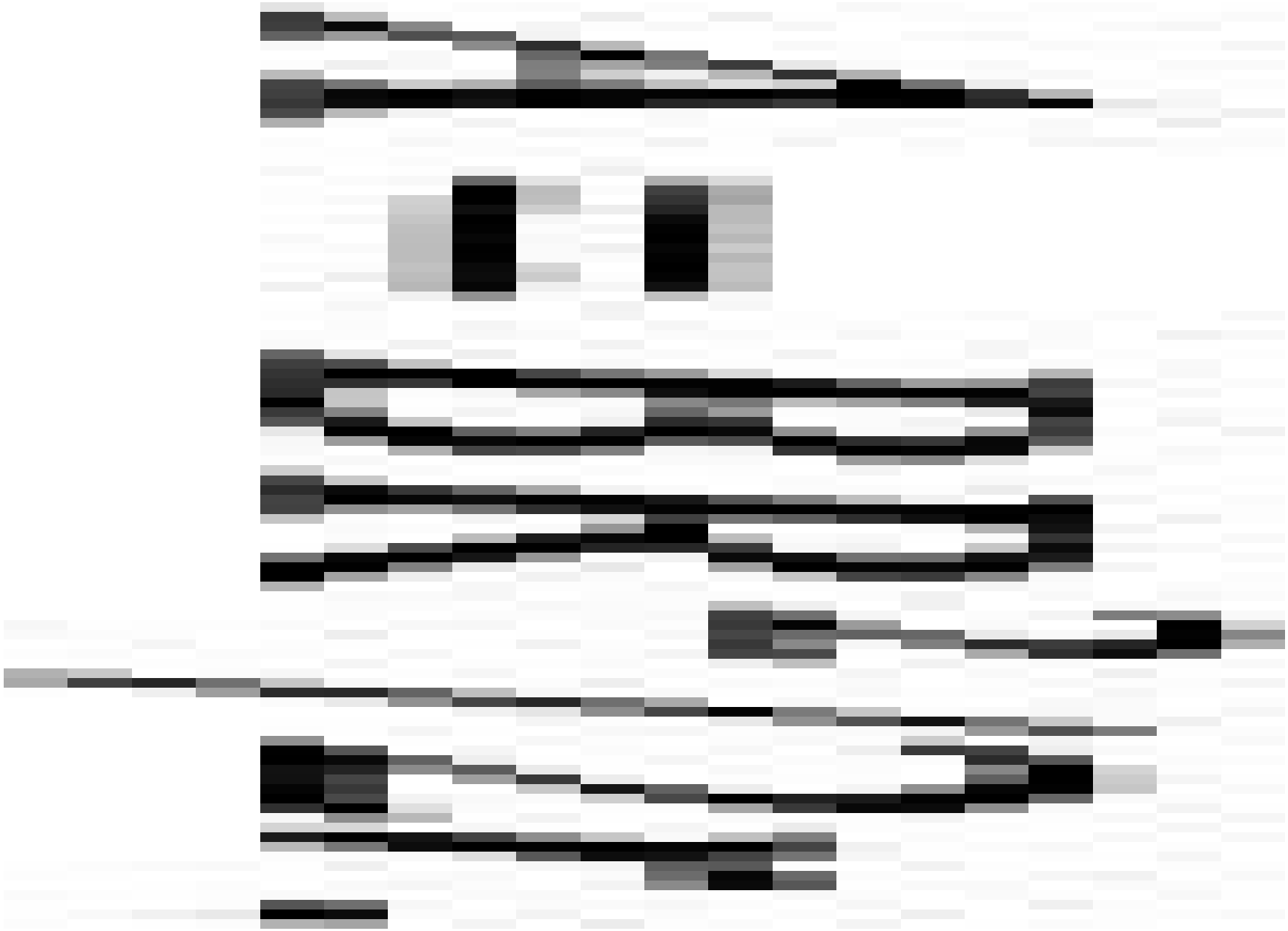
and hence



Around a similar circle outside the solenoid



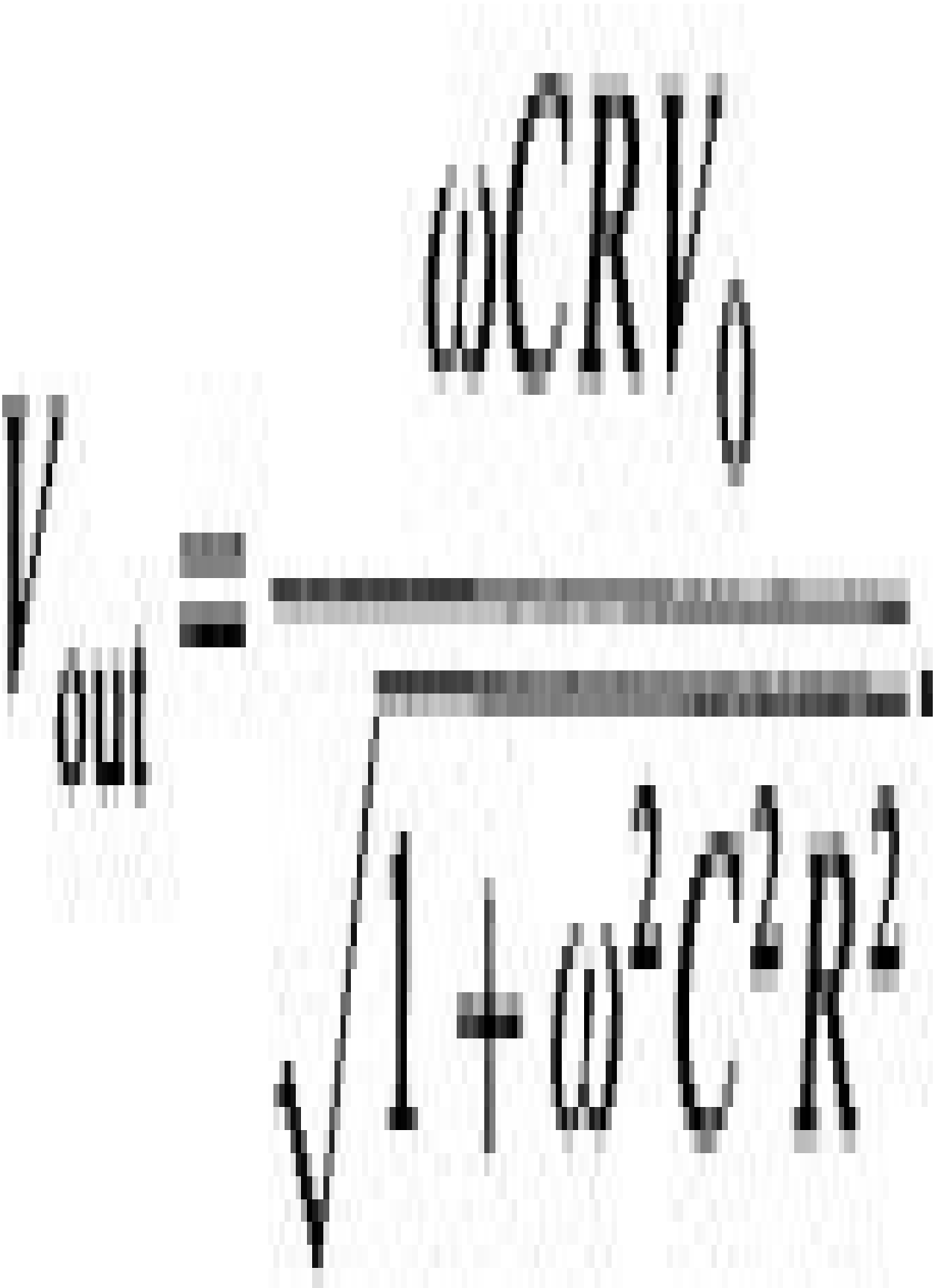
and



The solution for the potential A is analogous to the solution for the field B inside and outside a long straight wire carrying a uniformly distributed current. The field B depends only on perpendicular distance from the wire and is tangential to circles drawn coaxial with the wire. Inside B is proportional to the distance from the wire; outside it falls off inversely proportional to distance.

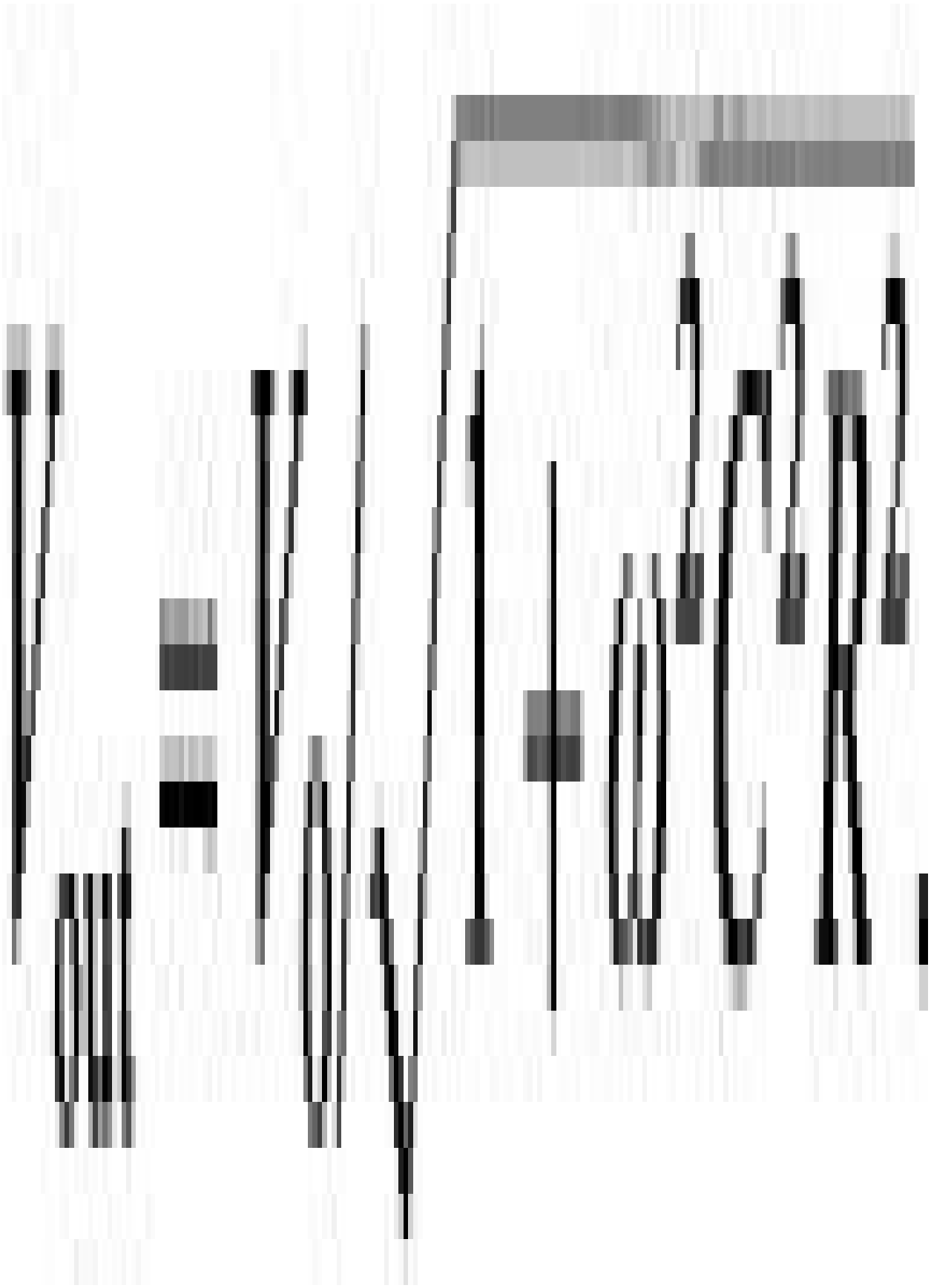
PROBLEMS 7

7.1 (a) The amplitude



Vout leads V0 by the phase angle $\phi = \tan^{-1}(1/\omega CR)$.

(b) The amplitude



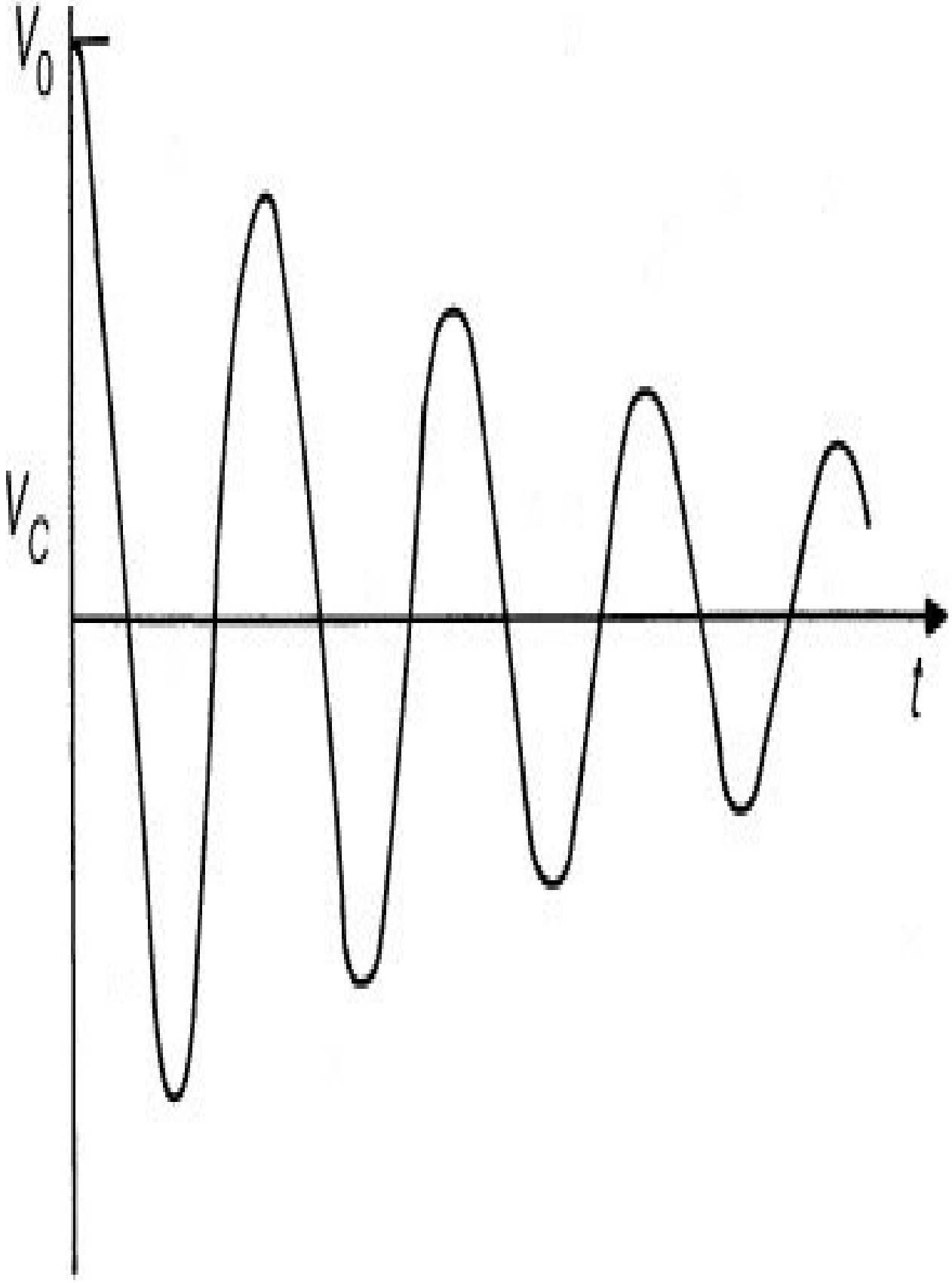
Vout lags behind V0 by the phase angle $\phi = \tan^{-1}(\omega CR)$.

7.3 The power factor is 3.2×10^{-3}.

7.6 A 4 μF capacitor gives critical damping: $Q = \omega_0 L/R =$ for any critically damped circuit

7.7 The natural angular frequency is . This differs by only about one part in 10^3 from the resonant angular frequency . The logarithmic decrement $\pi R/\omega_n L = \pi/10$, and the oscillation decays by $\exp(-\pi/10) \approx$ in each period. The initial condition is Vc = V0 and the final state when the transient has died away is V0 = 0. The resulting transient waveform is sketched in Figure D.4.

Figure D.4.


V_0
V_C
t

PROBLEMS 8

8.1 Ω.

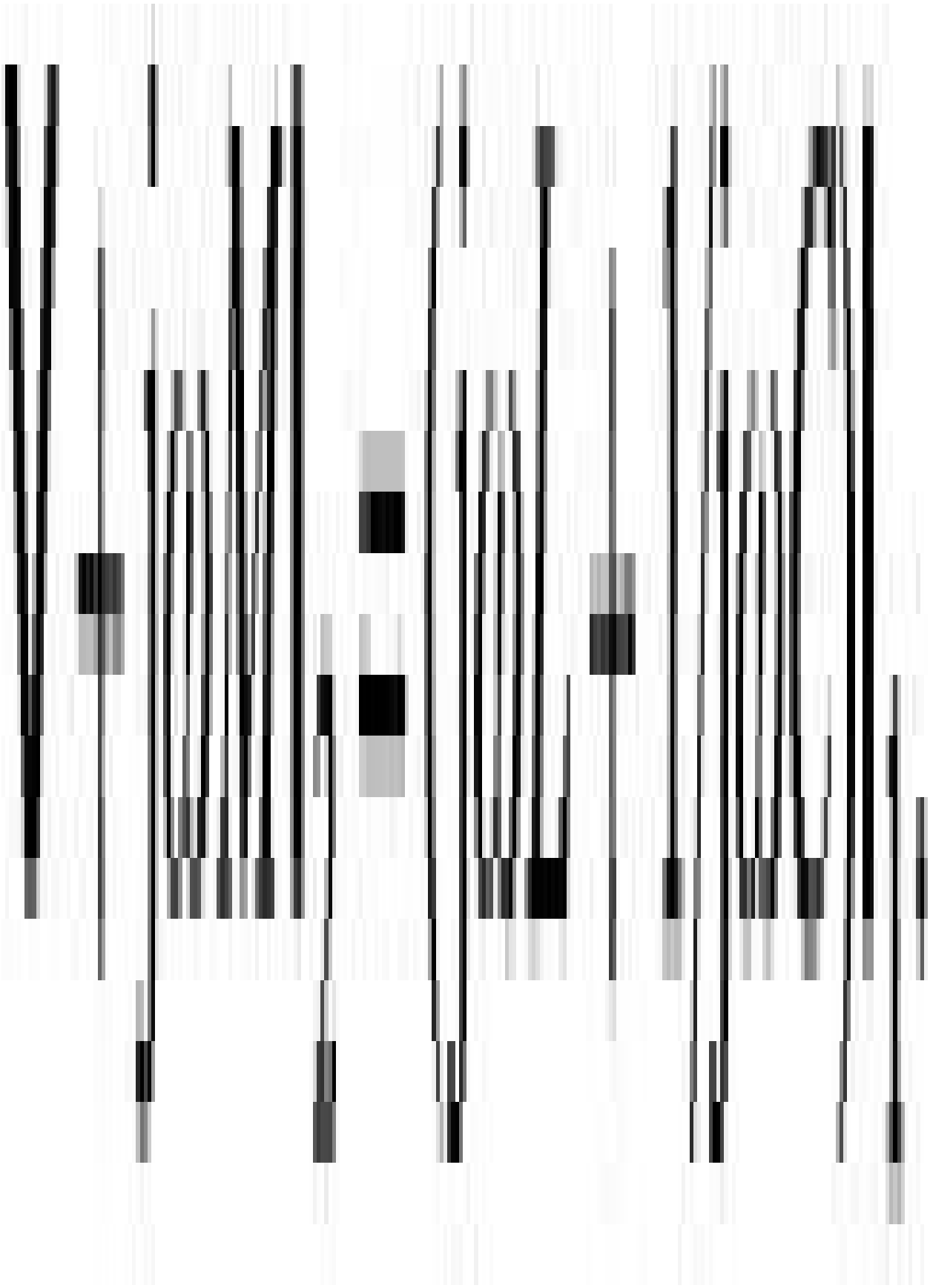
8.2 In the complex number notation, the loop currents are , and the voltage across C is $(I2 - I1)/j\omega C = V0$, i.e. a voltage in phase with the generator and with amplitude V0. This example can be solved more easily by using the principle of superposition, considering the currents driven by each generator separately. Each sees an impedance $3/2j\omega C$, made up of a series capacitor and two in parallel, when the other generator is removed.

8.3 The admittance of the parallel resonance circuit is $Y = 1/(R + j\omega L) + j\omega C$. This can be rearranged to give $Y = (RC/L) + j(\omega C - 1/\omega L)/(1 - jR/\omega L) \approx (C/L)[R + j(\omega L - 1/\omega C)]$, neglecting the term $jR/\omega L$. As $I = VY$, the current falls to a minimum value near the angular frequency $\omega 0 = 1/$. The voltage across the capacitor is V0, so that the energy stored by the circuit is . The impedance at $\omega 0$ is L/CR, leading to an energy dissipated in one radian , and Q =(stored energy)/(dissipated energy) = $\omega 0L/R$ as for the series circuit. The neglect of the term $jR/\omega L$ is thus justified for a high-Q circuit, i.e. if $L \; CR^2$.

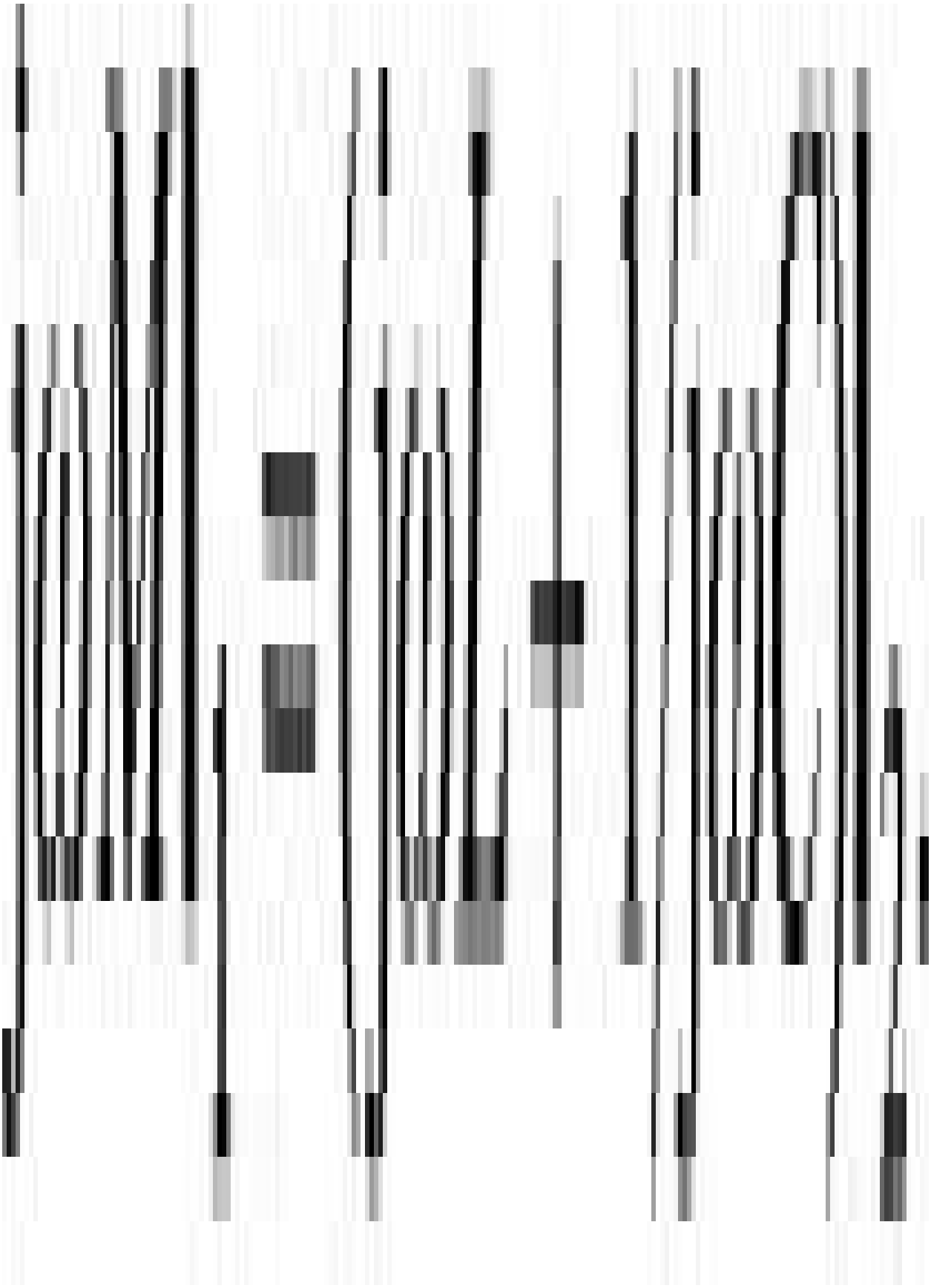
The part of the current in phase with the voltage is $V0CR/L$, and the net power delivered by the generator near the resonance is . Note that if the voltage generator is replaced by a constant current source, provided that we may still neglect the term $jR/\omega L$ the voltage has the same variation with angular frequency as the current in the series circuit (Equation (7.23)), and in this case the power exhibits a resonance with width $\delta\omega = \omega 0/Q$.

8.5 $L = CR1R2$, $R = R1R2/R3$. As with the Owen bridge, the balance conditions are independent of frequency and the variable quantities C and R3 each appear in only one of the balance equations.

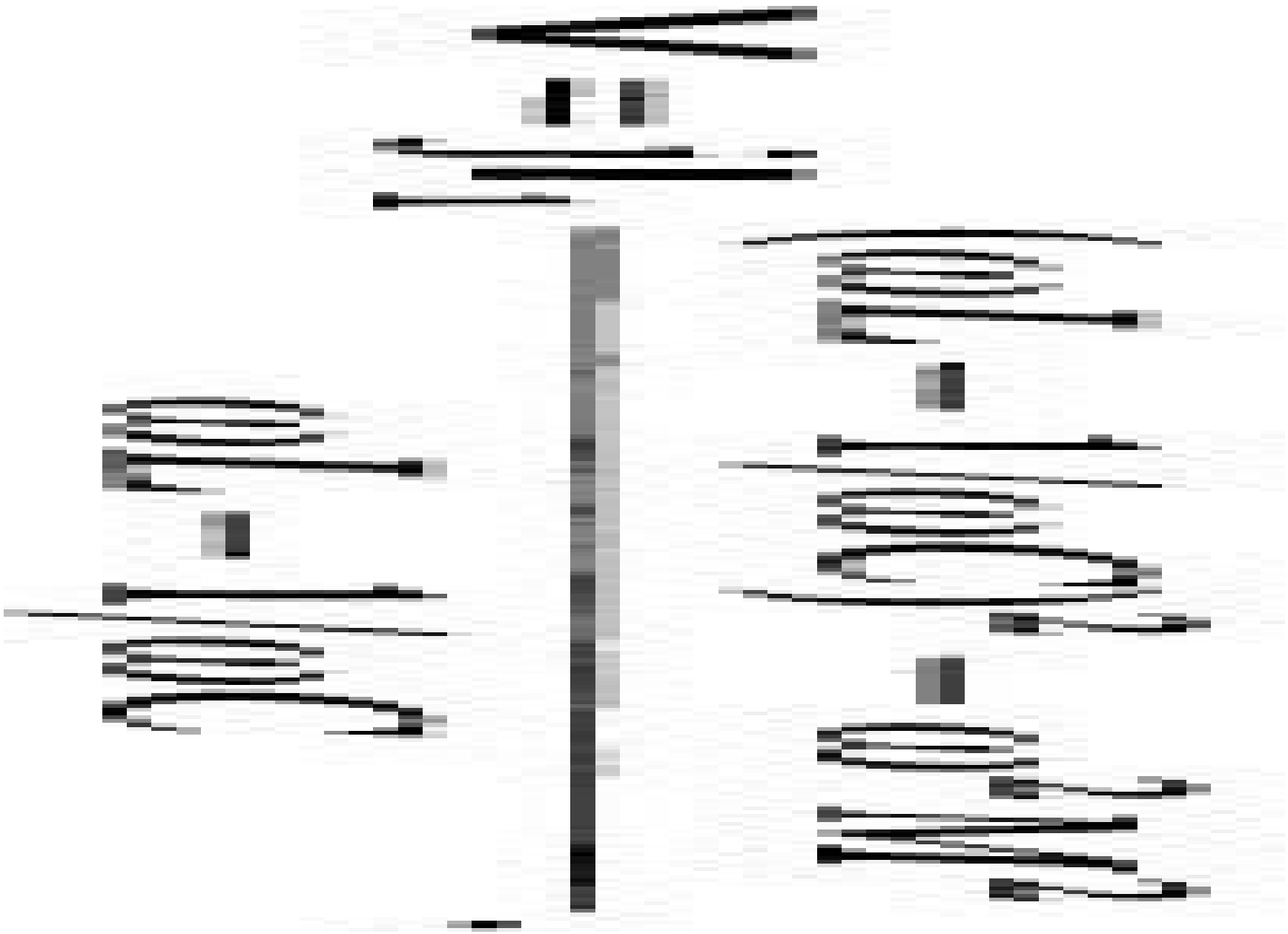
8.7 The current I1 in loop 1 induces an e.m.f. $MdI1/dt = j\omega MI1$ in loop 2. The current I2 in turn induces an e.m.f. $j\omega MI2$ in loop 1. The loop equations are thus



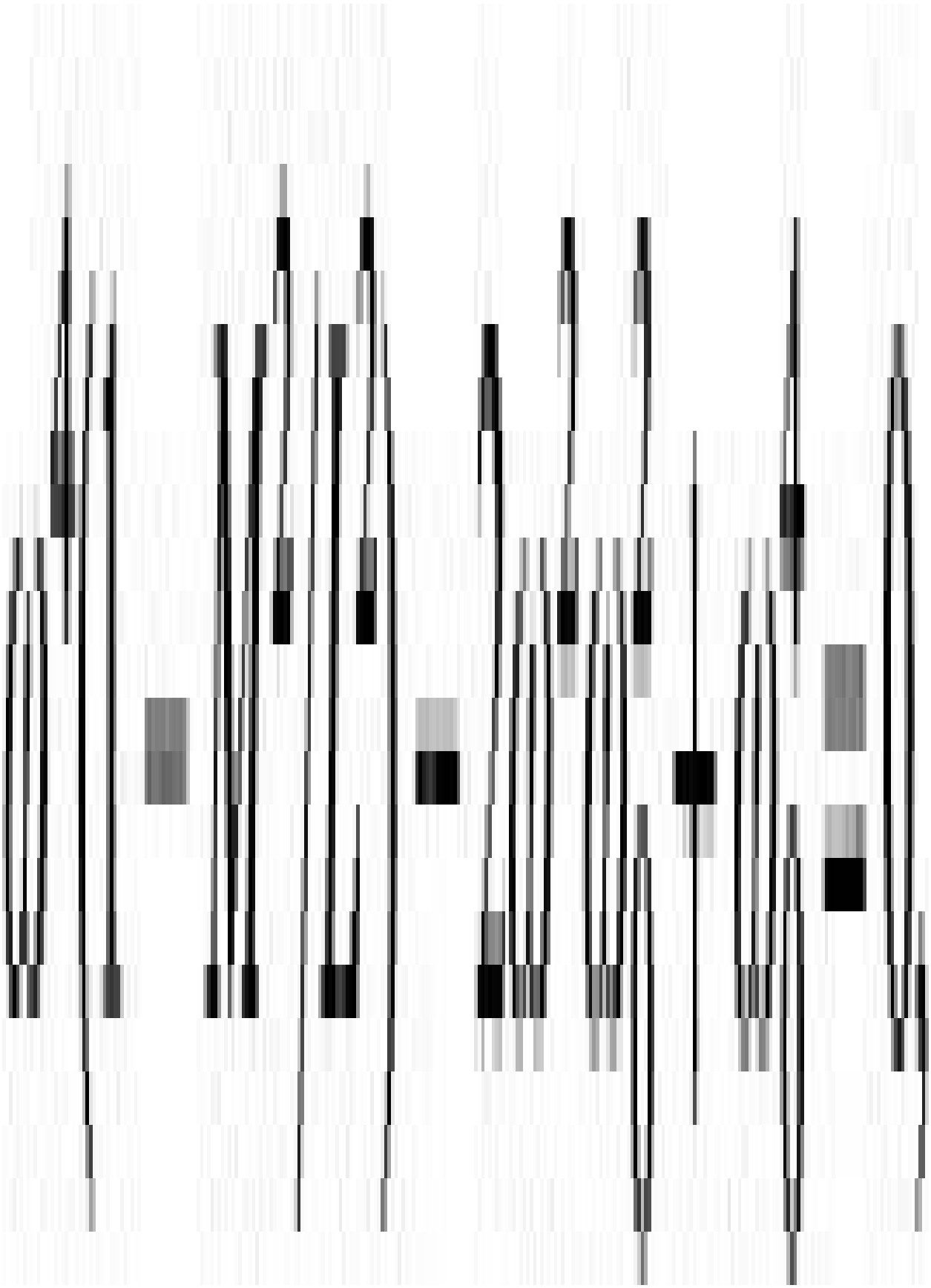
and



Eliminating I2 leads to

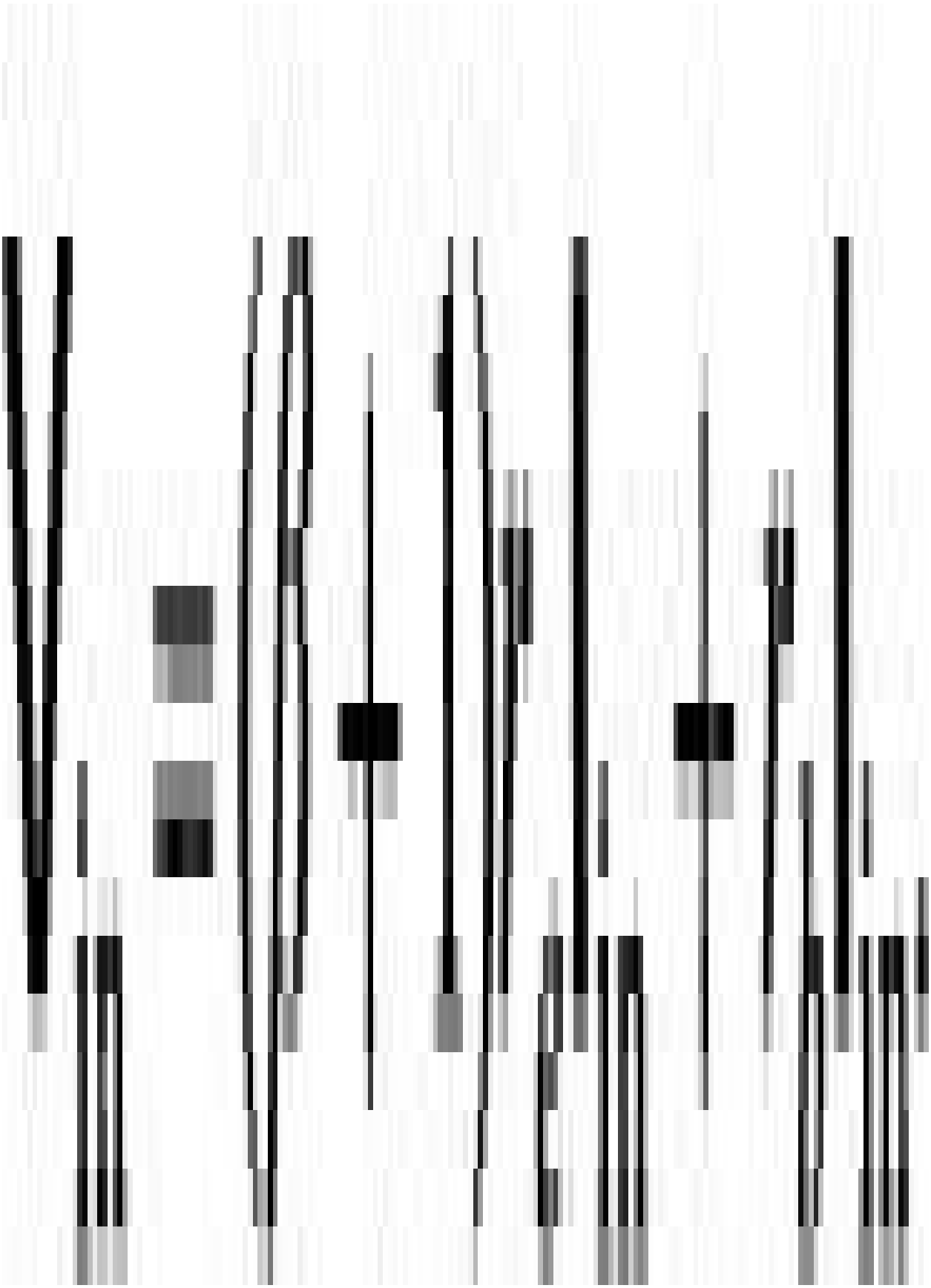


The impedance presented to V is zero if $(\omega L - 1/\omega C)^2 - \omega^2 M^2 = 0$. Introducing ( ), this may be rewritten as

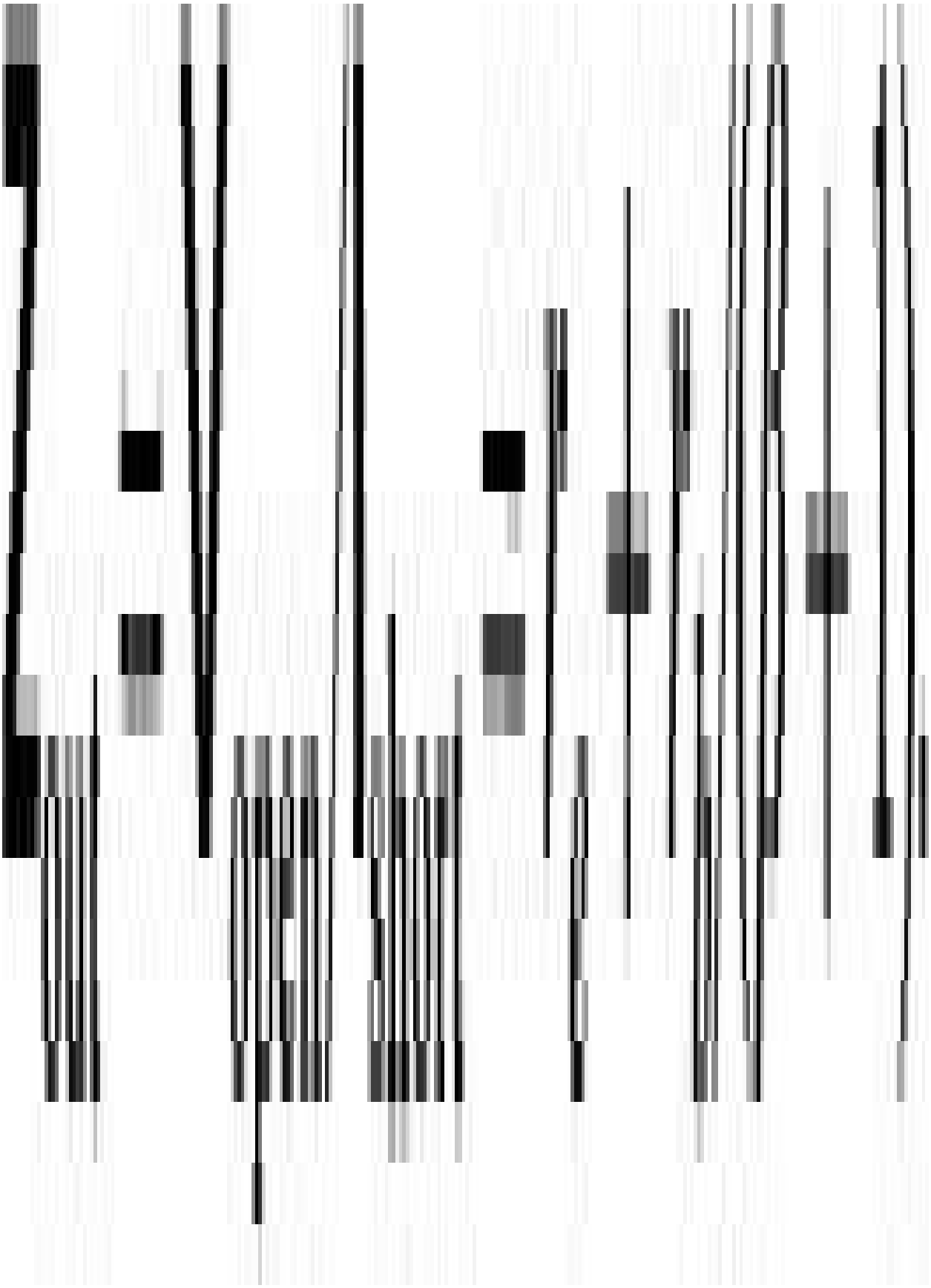


leading to . If the idealized self-inductances are replaced by coils which have a D.C. resistance the resonant frequencies are shifted by a very small amount if $\omega L/R$ is large and if M/L is not close to 1.

8.9 The open-circuit output voltage is $V_{open} = V_{in}$. If the load is short-circuited, the current delivered is $I_{short} = (\beta + 1)I_{in}$, and the input voltage is

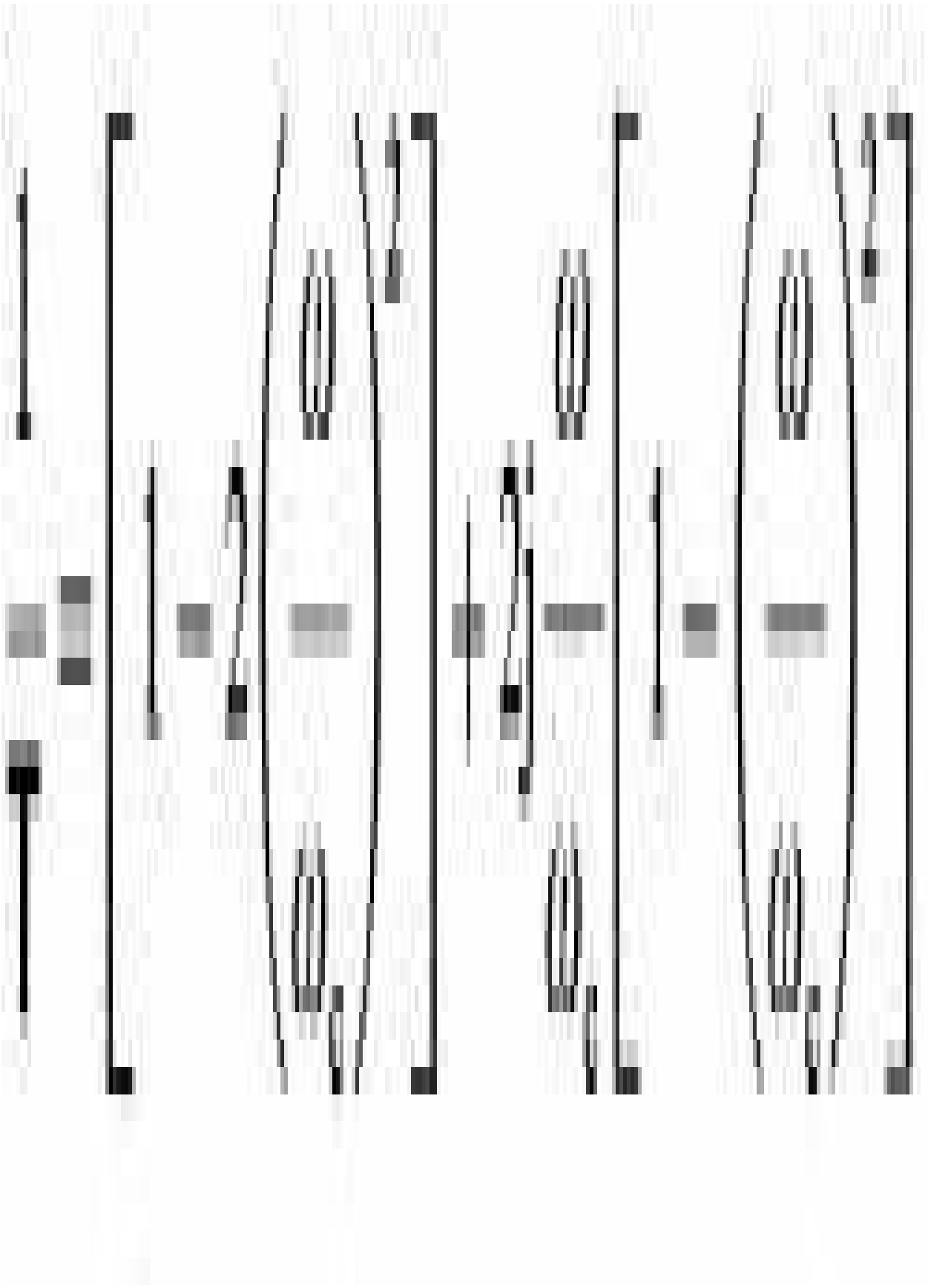


Hence

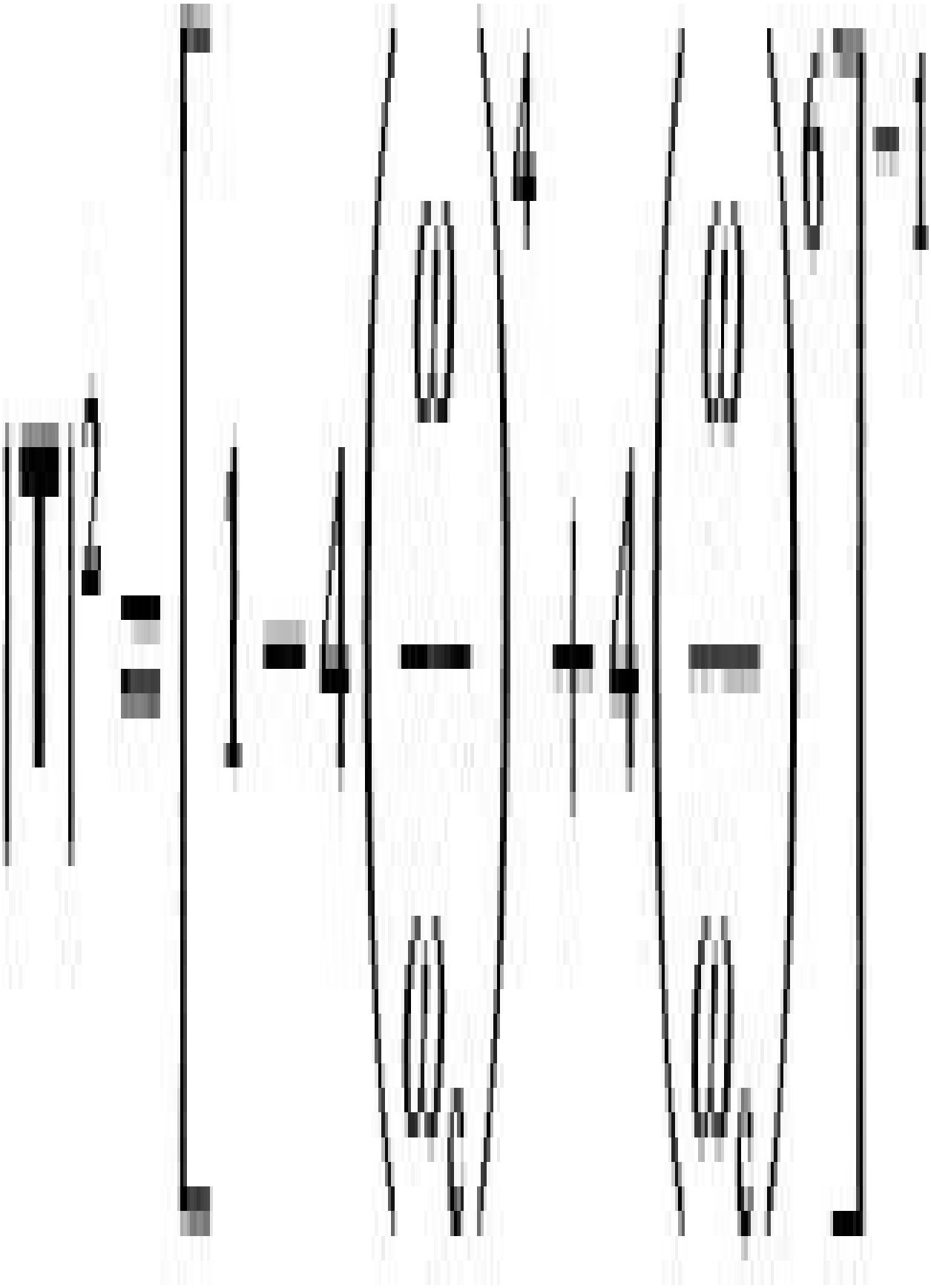


For a junction transistor re is small and the current gain β is large, so that Zout is small. The emitter follower is therefore suitable for an output stage, since the voltage it delivers is not sensitive to changes in the value of the load impedance.

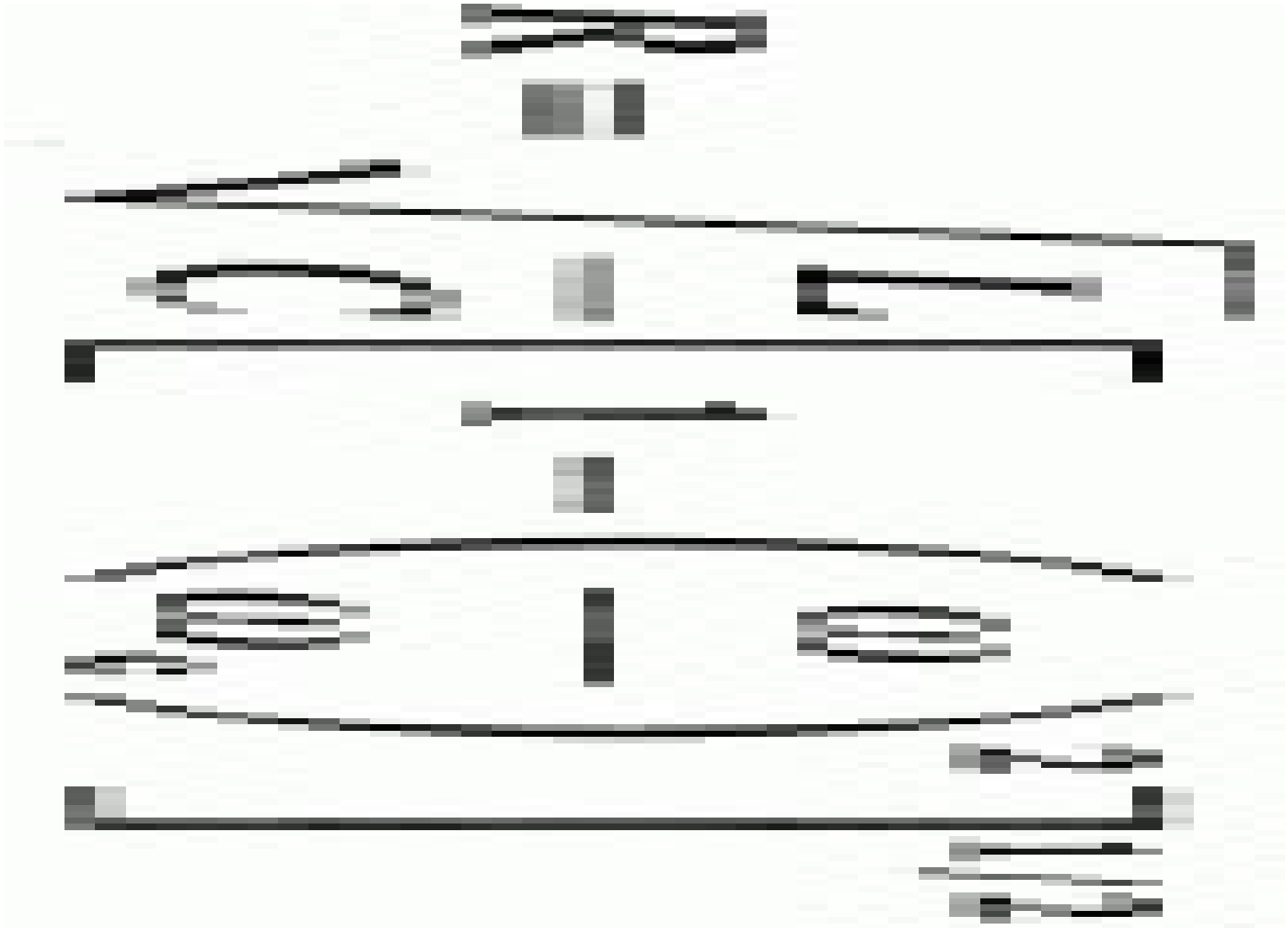
8.11 The voltage transfer function T is given by



where ωc = 2/ is the cut-off angular frequency. Hence



For frequencies up to (ω/ωc) = , T differs from unity by less than 3%, and low-frequency signals are less distorted by the T-section than by the L-section described in section 8.4. The reason for this is that at frequencies below cut-off the matching impedance for the T-section is a pure resistance



which differs from the terminating resistance of the network only to the second order in (ω/ωc).

8.13 The coupling coefficient k = 0.95.

PROBLEMS 9

9.1 Since the amplitudes of the incident and reflected waves are equal, the terminating impedance is purely reactive. For a node 75 cm from the end of the line, the reflection coefficient K = exp (−Jπ/2), requiring Z = −jZ0. The impedance −jZ0 corresponds to a capacitance of 16pF.

9.2 Ω.

9.3 The input impedance of the line is inductive, and the source of e.m.f. therefore sees a parallel resonance circuit with . The resonant condition is L = 1/ ω²C i.e. C = (ωZ0)−1 = 300pF.

9.4 The real part of the input admittance is (200 Ω)−1 one-sixth of a wave-length from the end of the line. The imaginary part can be cancelled by the input admittance of a short-circuited stub of length l, where tan kl = .

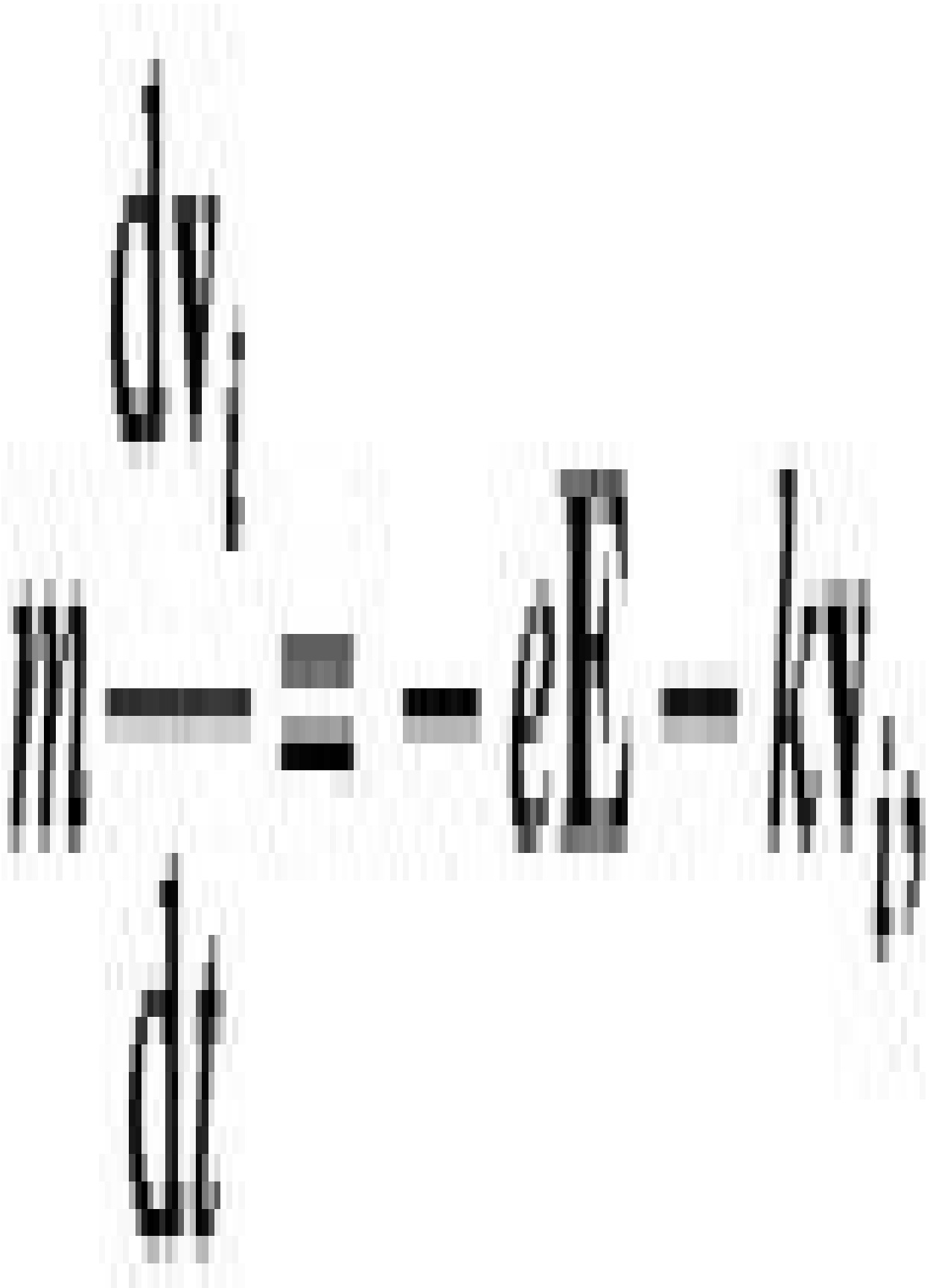
9.5 3/5.

9.6 Putting the reflection coefficient K = Kejδ, we find from Equations (9.20) and (9.21) that K = , δ = π/3, or . The impedance at the end of the line is .

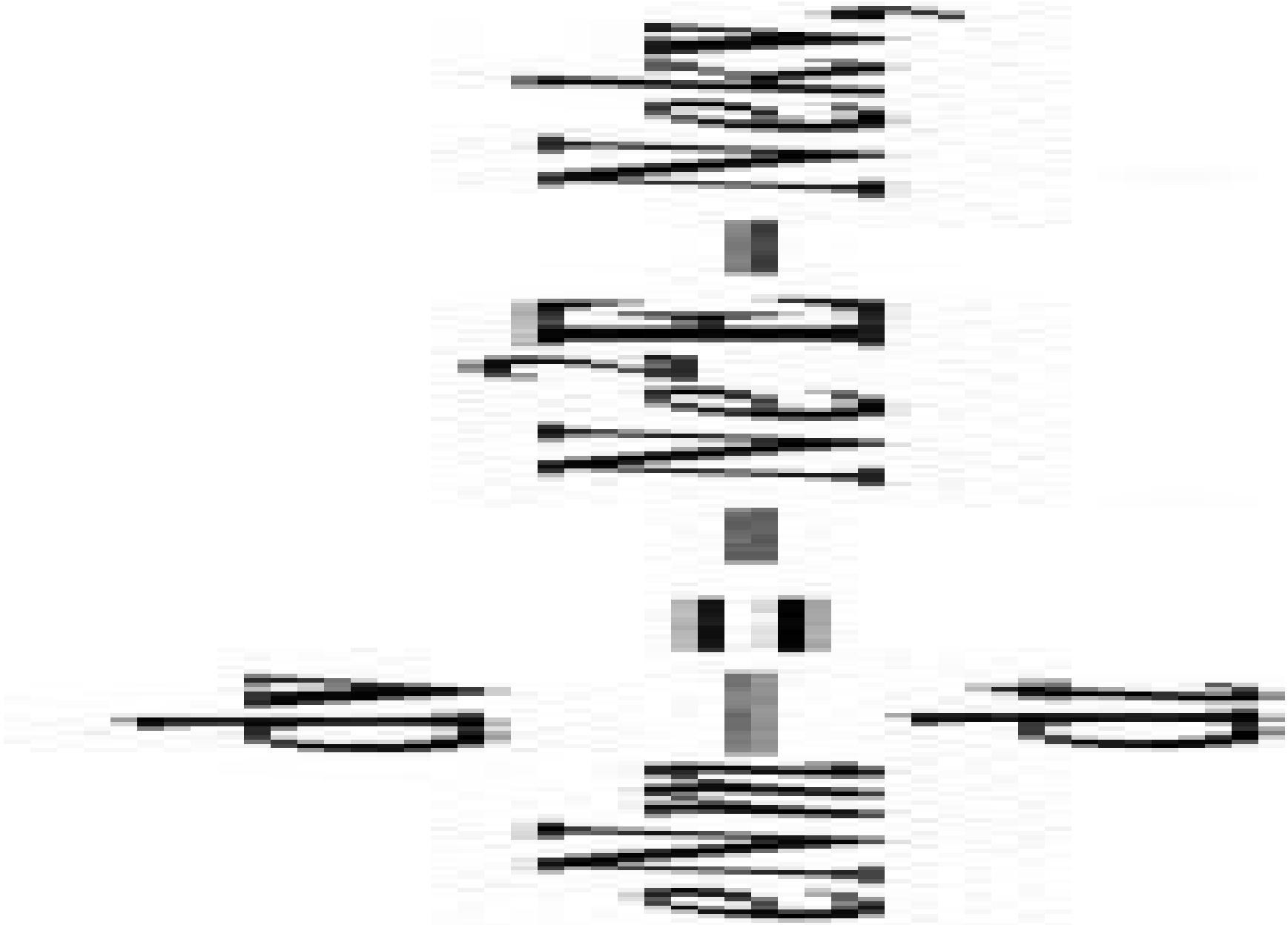
9.7 Signals leaving arm 1 in different directions reach arm 3 with a phase difference of π, and there is no output from this arm. The outputs from 2 and 4 are of equal magnitude but opposite phase. The hybrid ring thus allows high frequency signals to be divided without incurring power losses. With other choices for the impedances on the arms, a hybrid ring can also be used as a signal mixer, as a phase-changing or as a phase measuring device.

PROBLEMS 10

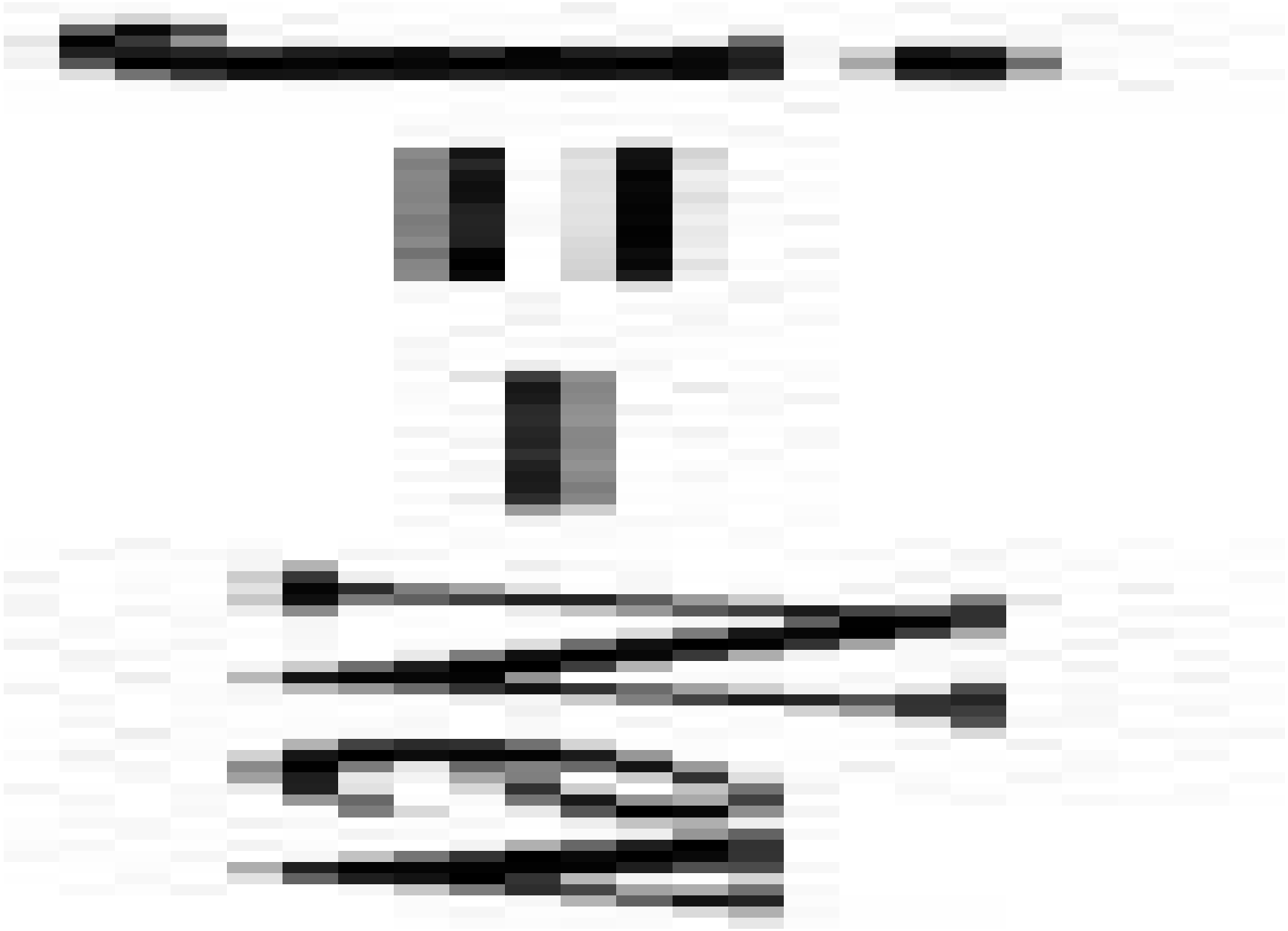
10.2 The conduction electrons have inertia and will not follow very rapidly changing fields. The relation $j = \sigma E$ breaks down at frequencies greater than about 10^{11} Hz. (Classical theory is not often useful at such high frequencies). In this problem it is best to begin with the equation of motion of an electron and include a frictional damping force to account for the retarding action of the surroundings. Hence



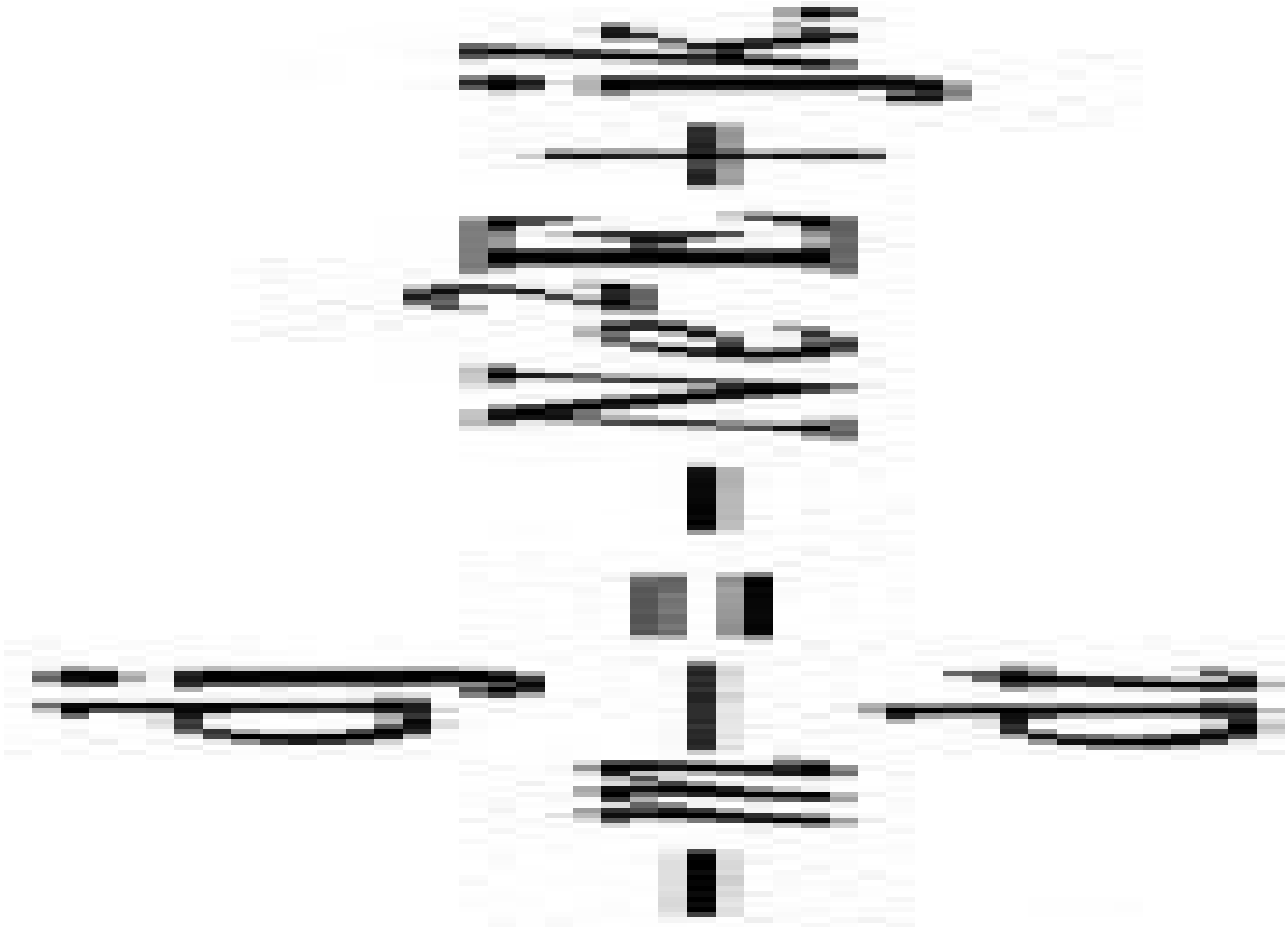
where the label i refers to a specific electron, and we use the symbol m for the electron mass. If we apply this equation to the N free electrons in unit volume of the metal and average, after multiplying by e we have,



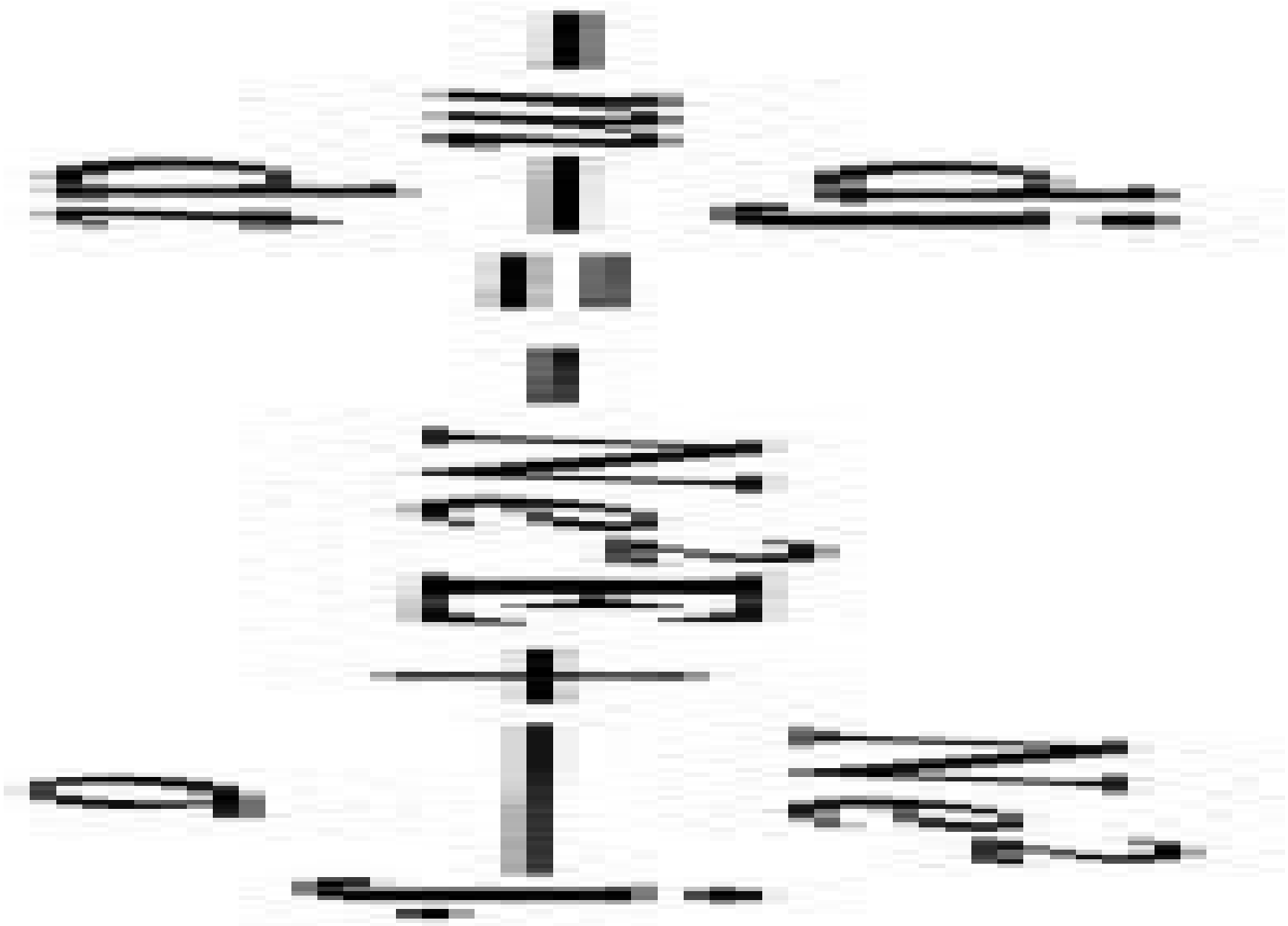
where v is now the mean drift velocity of the electrons in the direction of the field E. Now



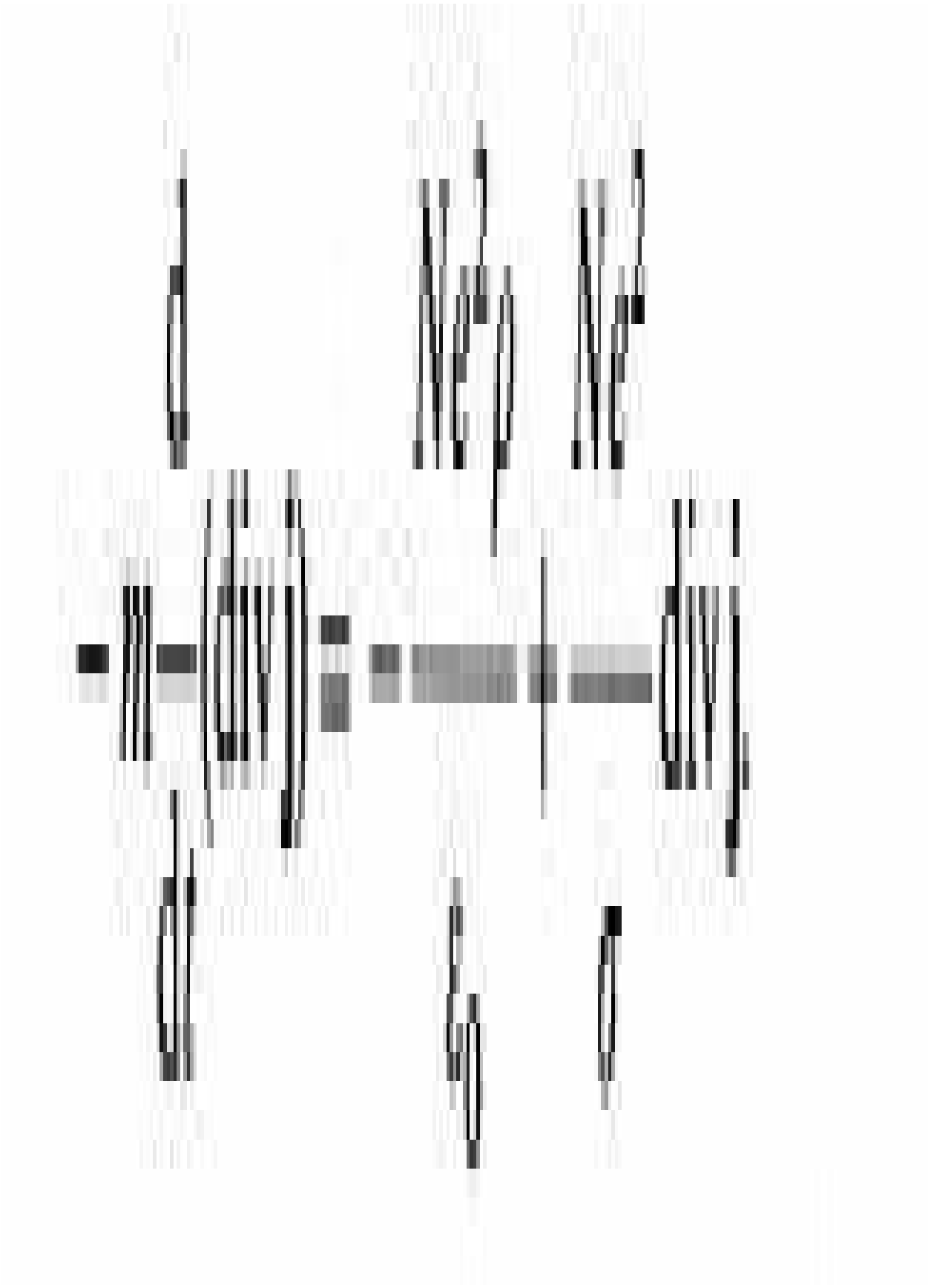
from Equation (4.1), and so



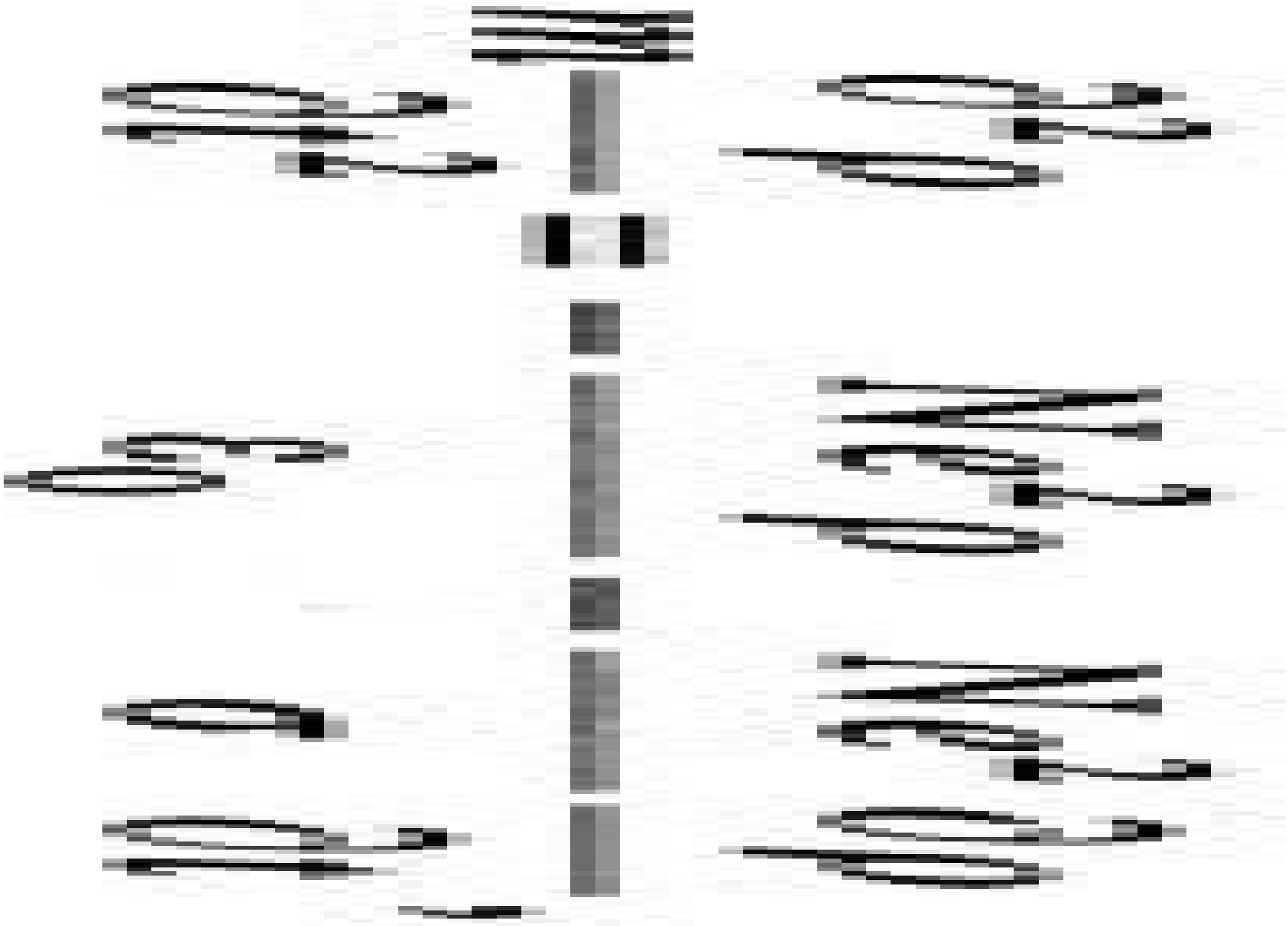
For steady currents $j = \sigma E$ and v is constant, giving $k = Ne^2/\sigma$. The above equation then becomes



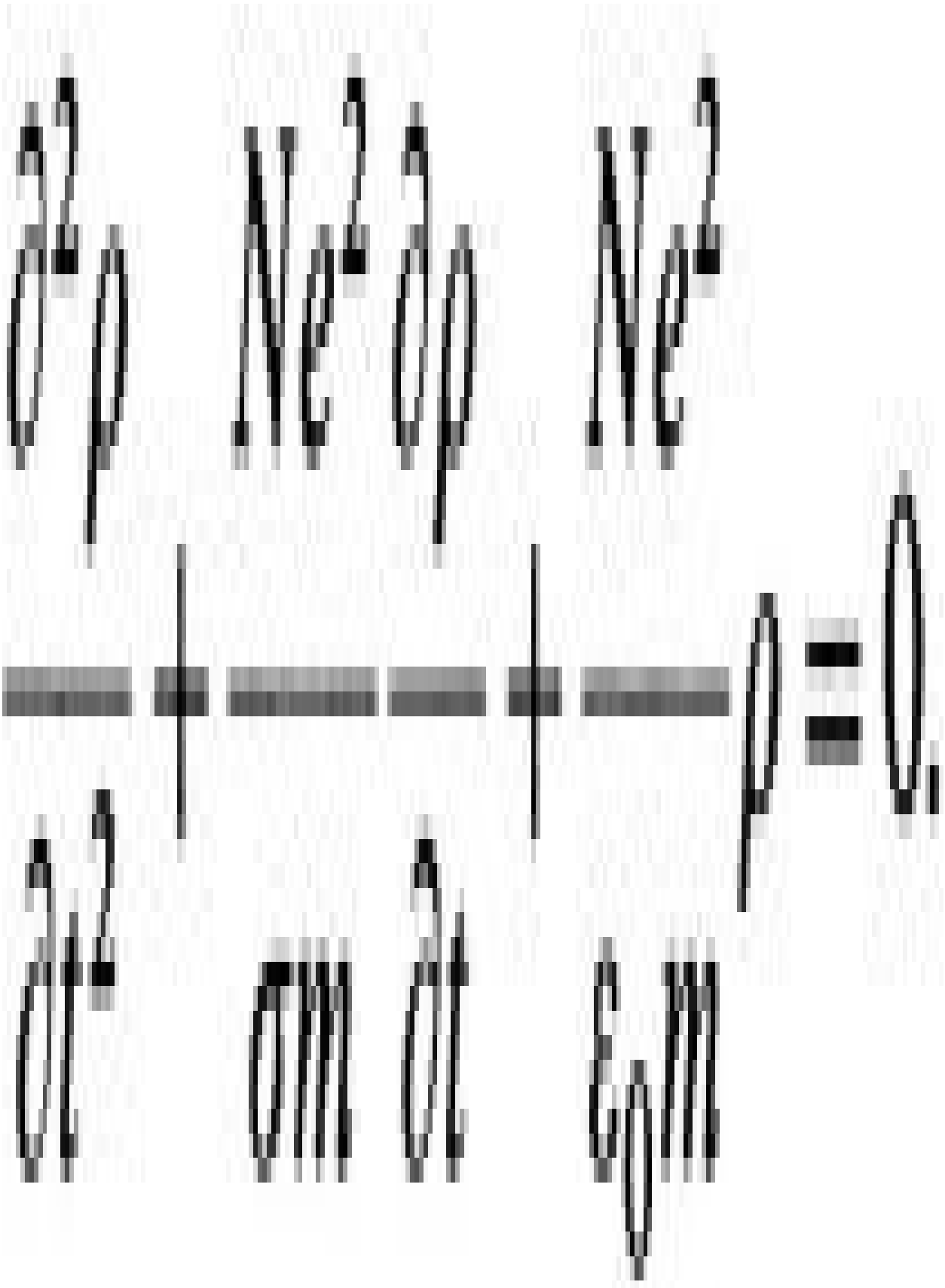
The field E can be replaced using Equation (1.20), div $E = \rho/\varepsilon 0$, and we obtain



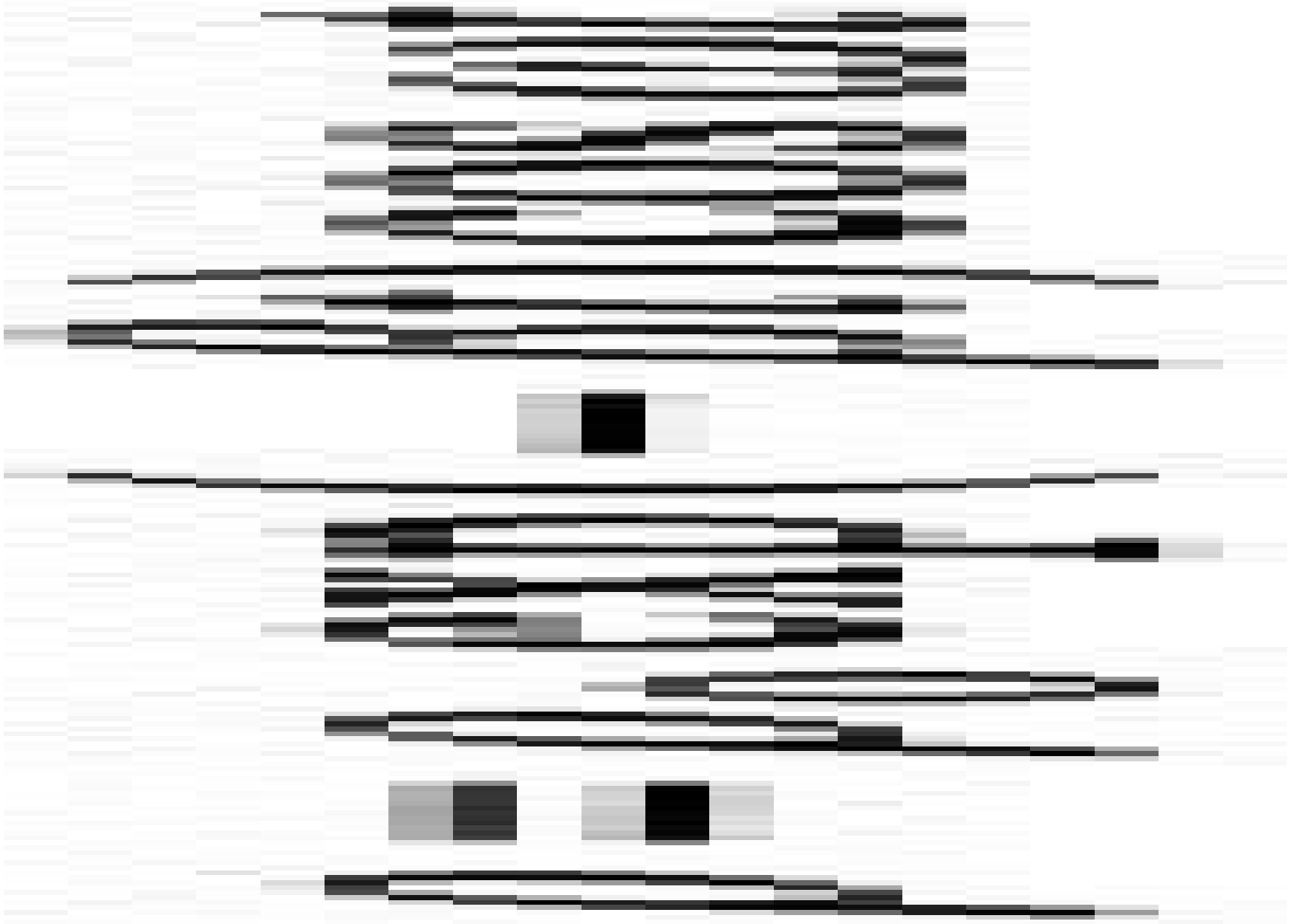
We have assumed that the relative permittivity of the metal is unity. Using the equation of continuity (10.1) this equation can be rewritten



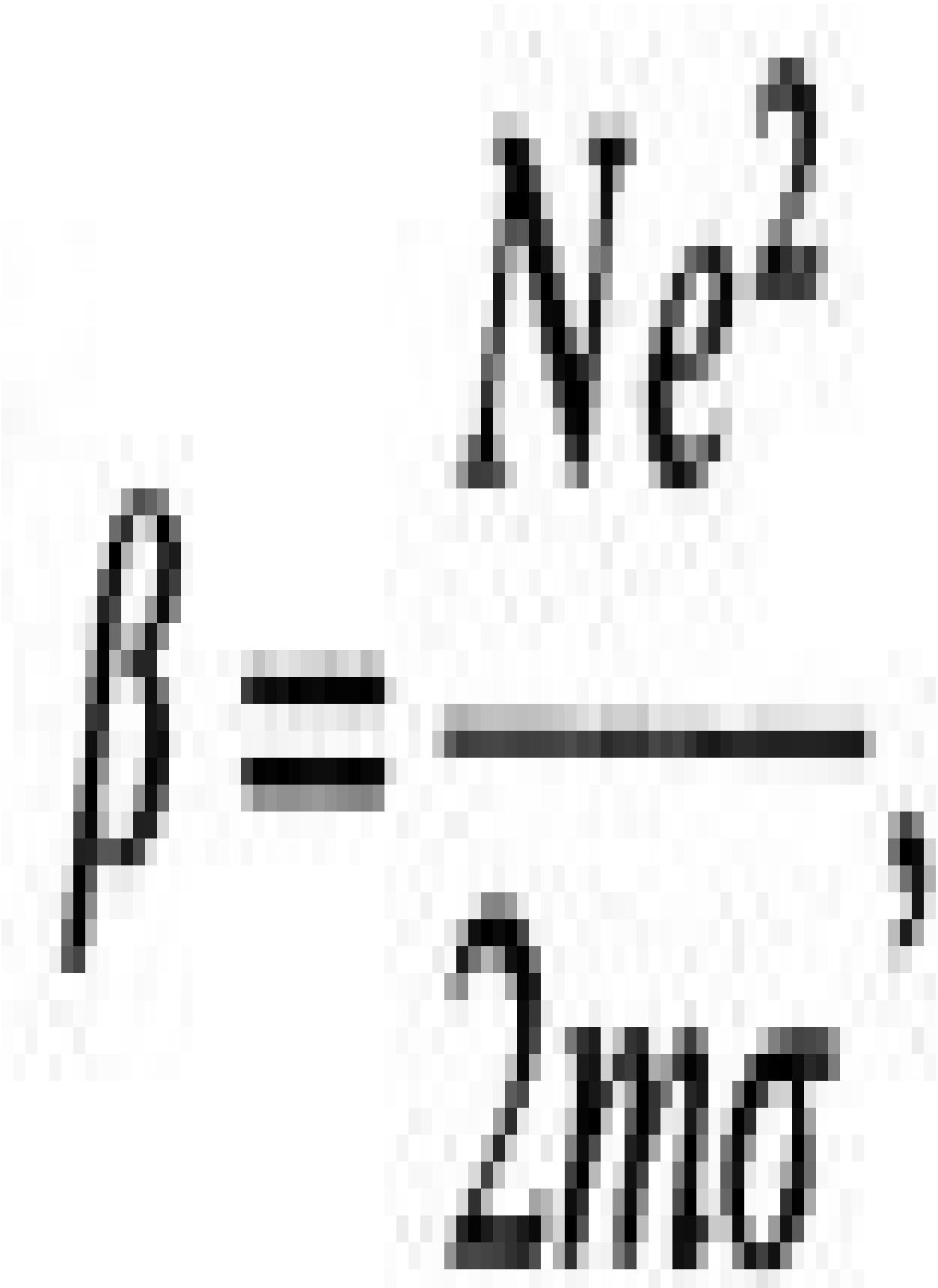
or



This equation tells us how the charge density ρ varies with time. It represents damped simple harmonic motion if $\varepsilon_0 Ne^2/2m\sigma^2 < 1$, which is true for metals. The solution is



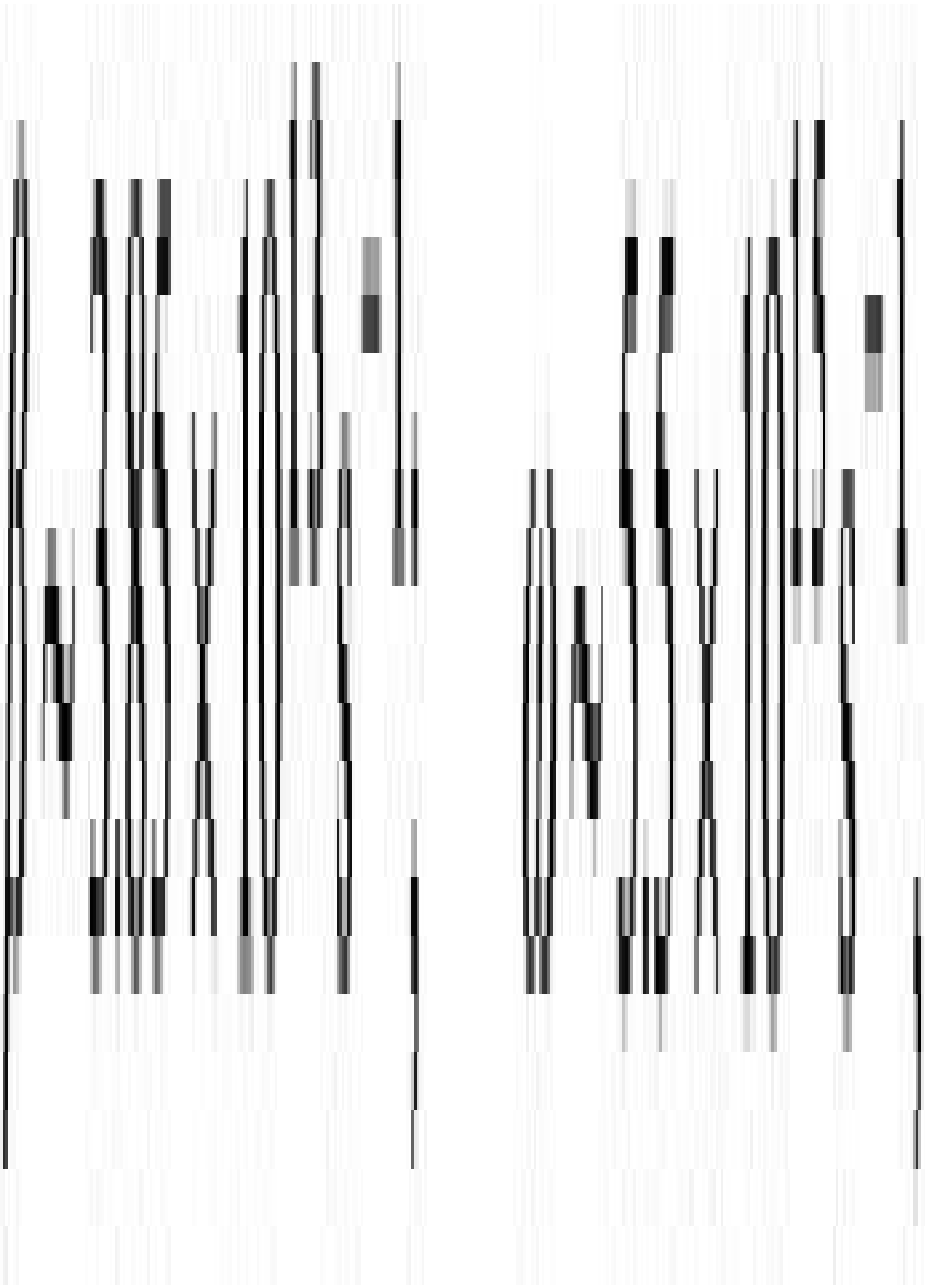
where



and ω, the plasma frequency, is approximately given by

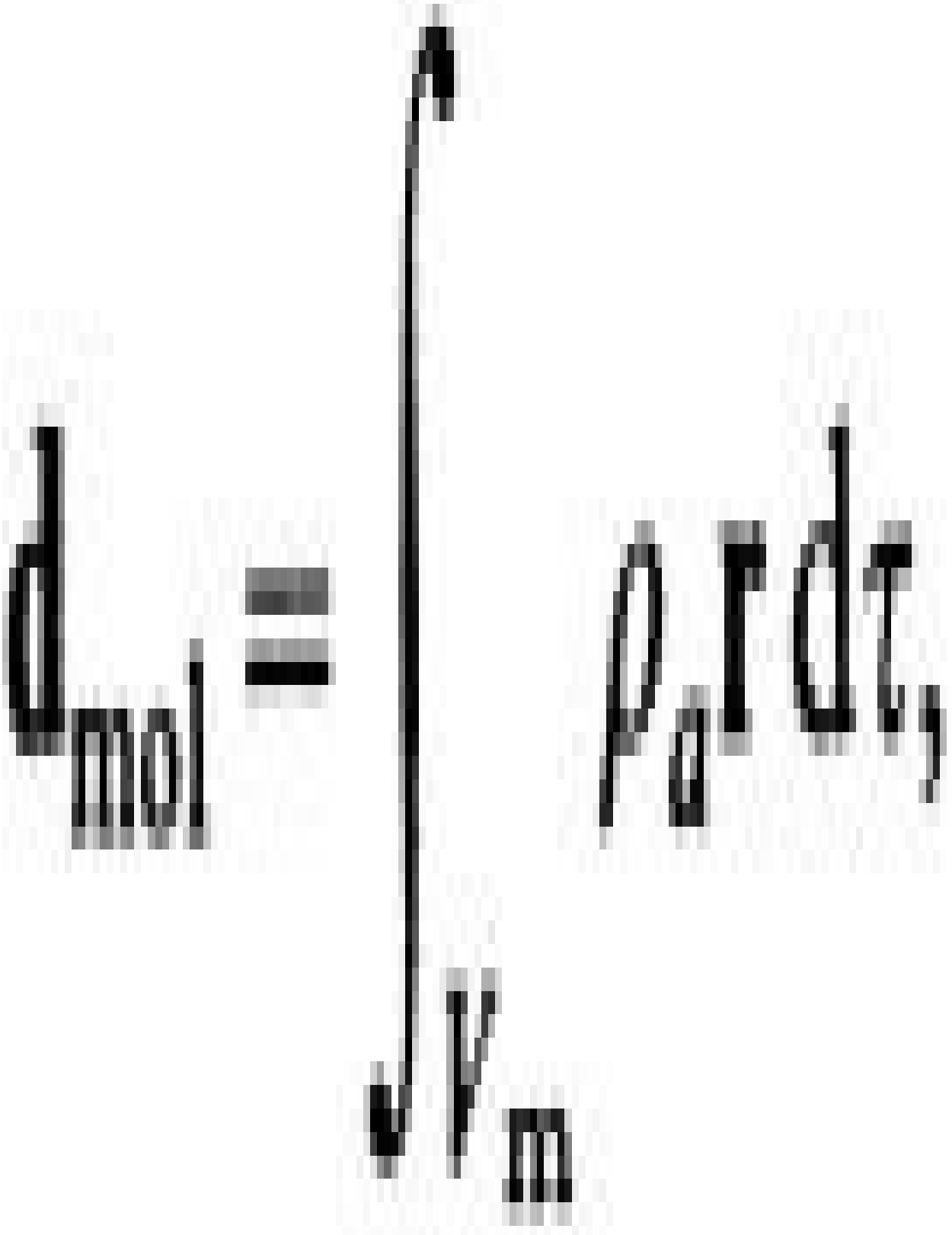
$$\omega^2 = \frac{Ne^2}{\varepsilon_0 m},$$

Substitution of the values for copper in these equations gives



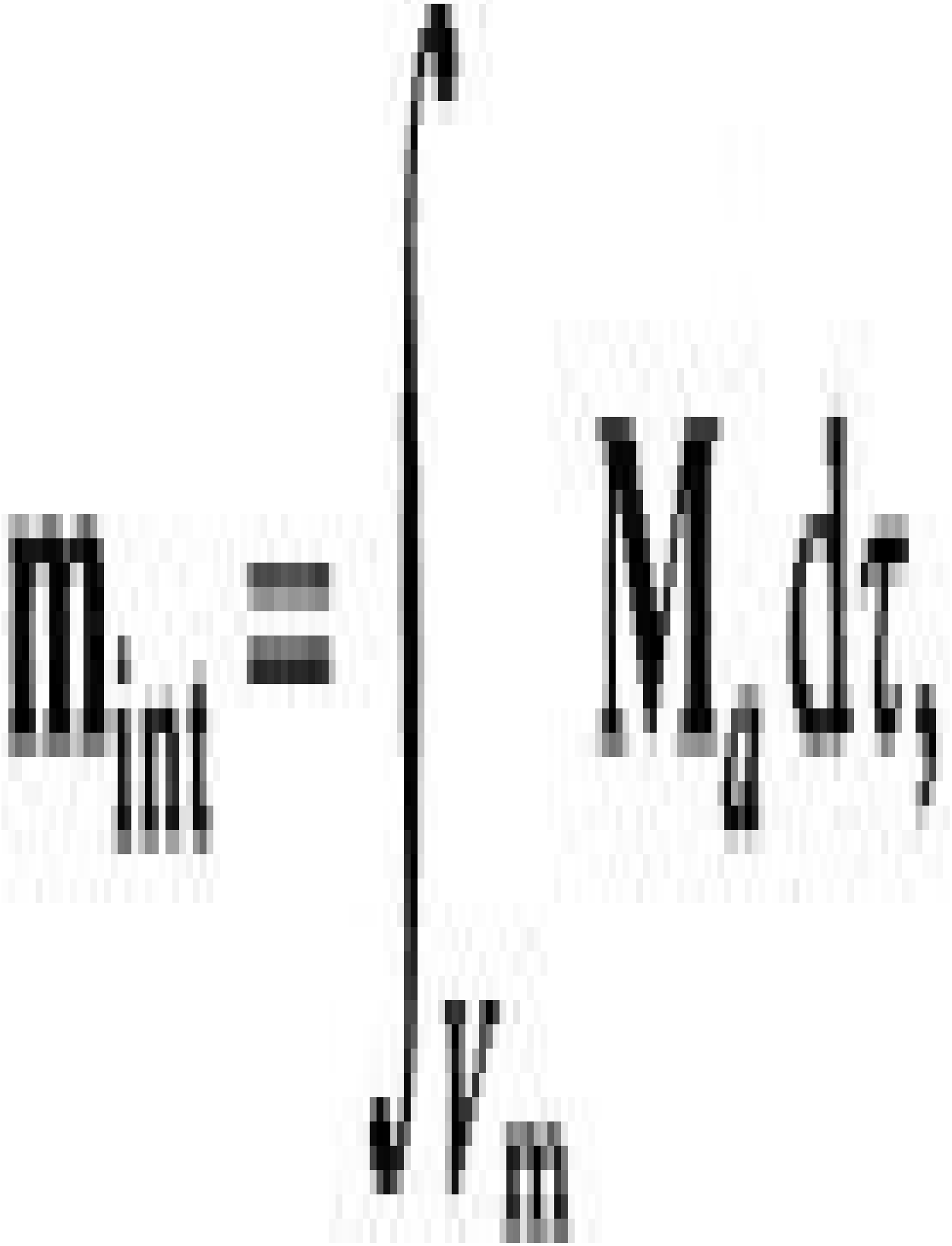
assuming that there is one free electron for each copper atom, and using the value of σ given in Table 4.1 (p. 115). The oscillations thus die away to one half of their amplitude in a time ~ 1.8×10^{-14} seconds.

10.3 The electric dipole moment is

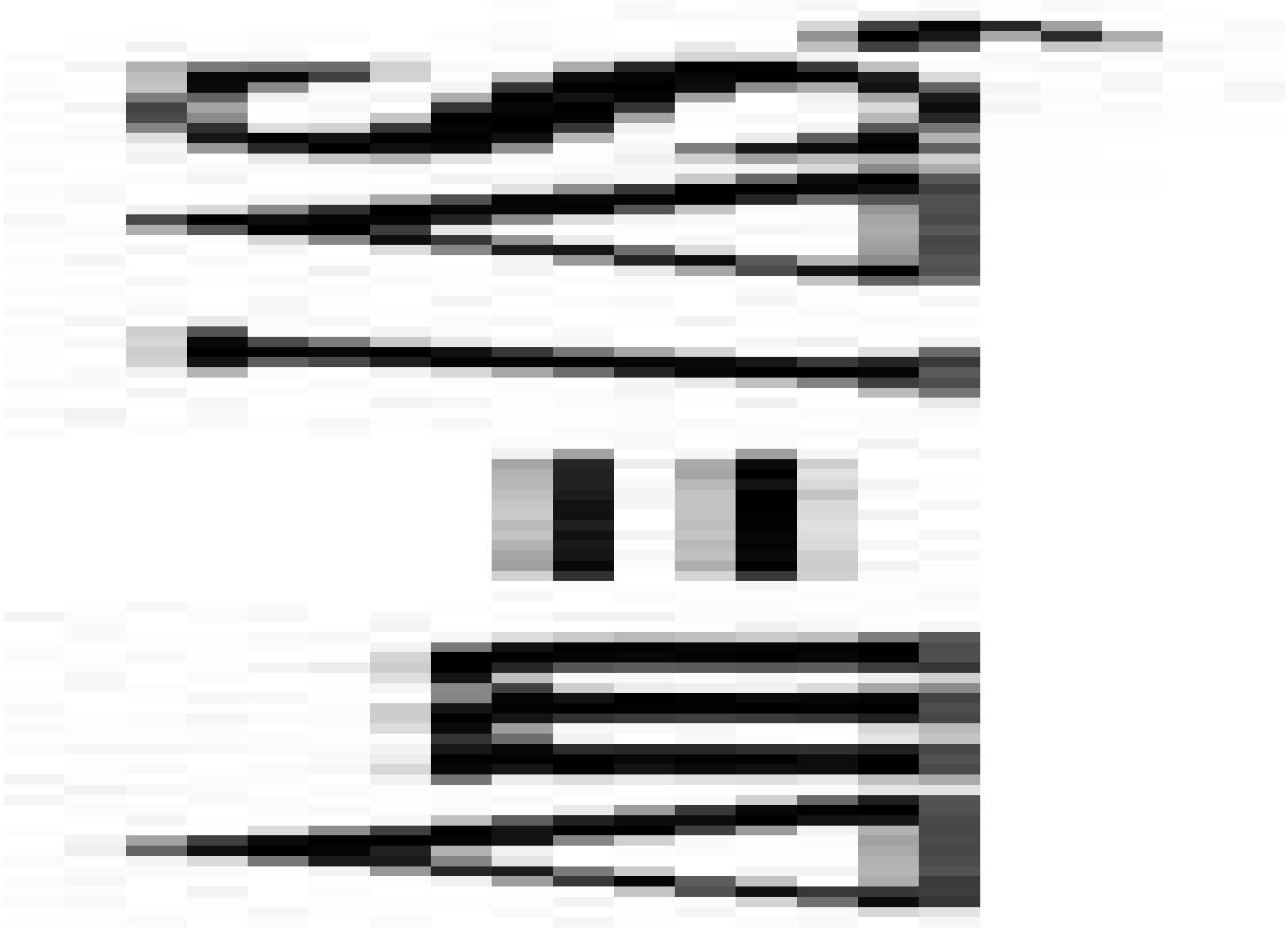


where Vm is the volume occupied by the molecule.

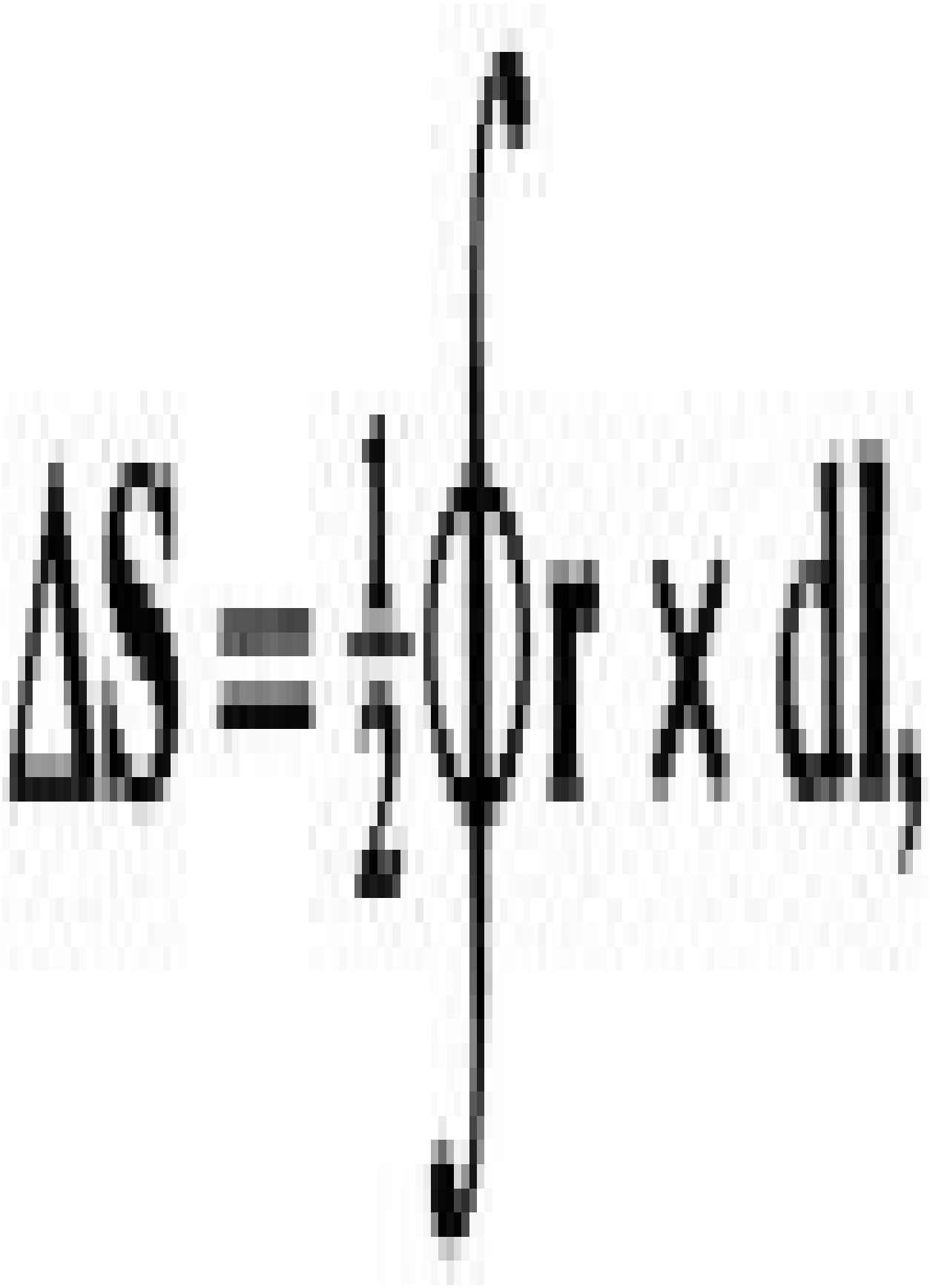
The magnetic dipole moment has two contributions, one from the intrinsic dipoles given by



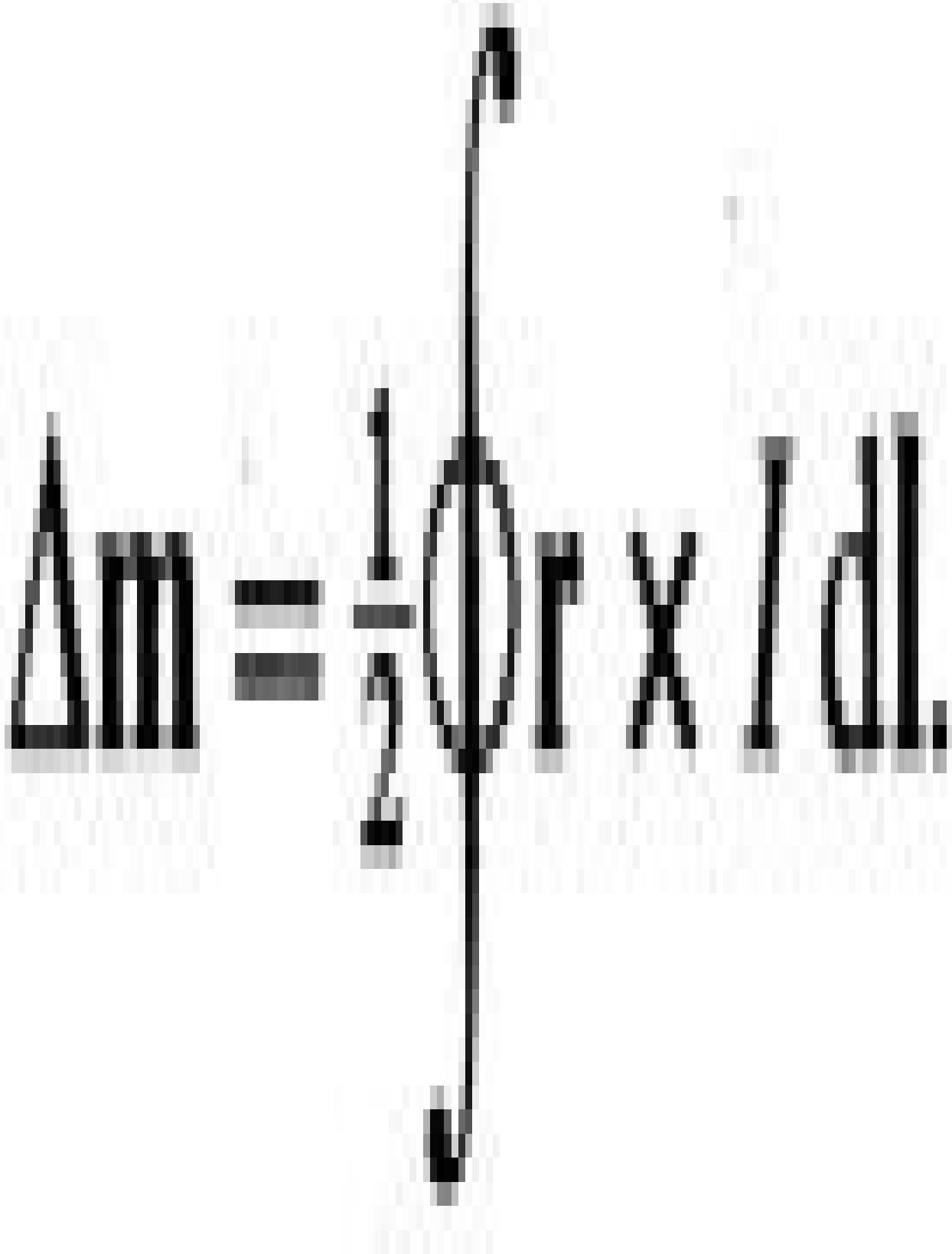
the other mcurrent from the distribution of currents within the molecule. The magnetic dipole moment of a current I in a small closed loop of area ΔS is given by



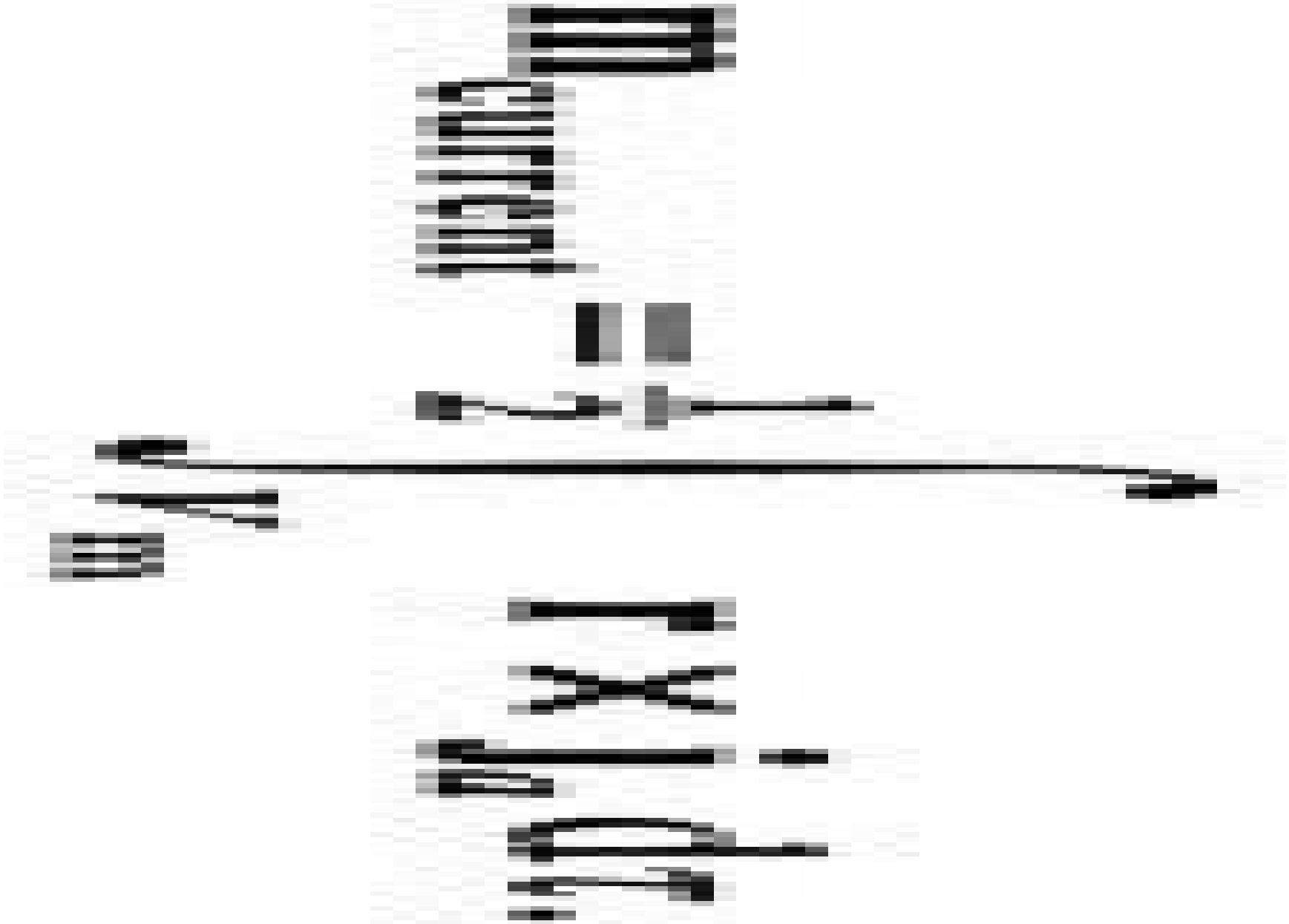
where ΔS is a vector normal to the plane or the loop, in a direction given by the right hand screw rule. It can be shown that


ΔS=∮r x dl

hence


Δm = 2πrxidL

For the current distributed over a volume Vm the vector I dl can be replaced by the vector ja dτ and



The total magnetic dipole moment of the molecule is thus

$$\mathbf{m}_{\mathrm{mol}} = \int_{V_{\mathrm{m}}} \left(\mathbf{M}_a + \tfrac{1}{2}\mathbf{r} \times \mathbf{j}_a\right) d\tau$$

$$= \int_{V_{\mathrm{m}}} \left(\mathbf{M}_a + \tfrac{1}{2}\mathbf{r} \times \mathbf{v}_a \rho_a\right) d\tau.$$

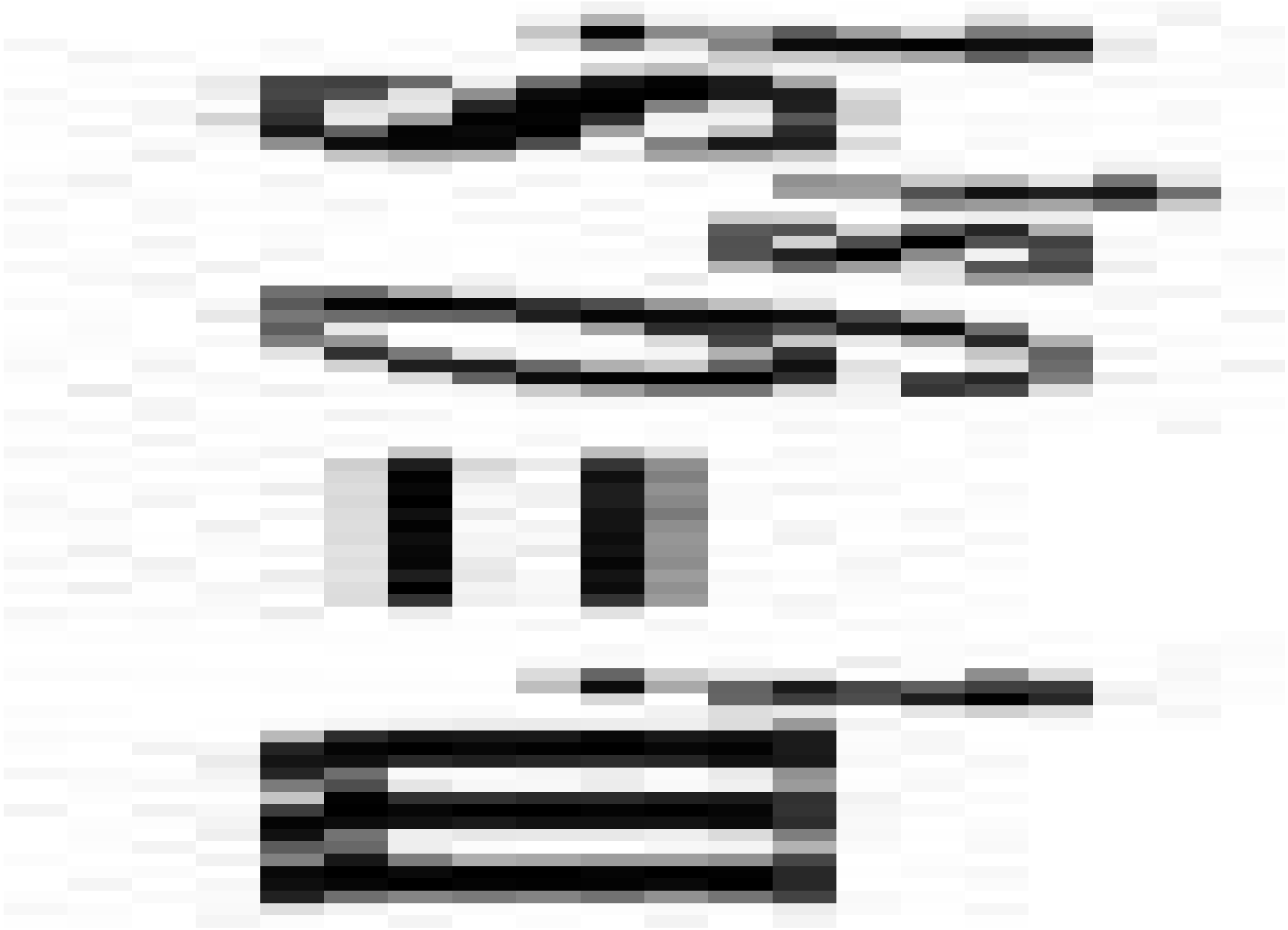
If the molecule is considered to consist of point particles the electric dipole moment becomes

where the sum is over all the particles in the molecule. The magnetic dipole moment becomes

$$m_{mol} = \sum_i \left(m_i + \frac{1}{2} r_i \times v_i e_i \right)$$

$$= \sum_i \left(m_i + \frac{e_i l_i}{2 m_i} \right),$$

where Ii is the orbital angular momentum of particle i. The intrinsic magnetic dipole moments mi are related to the intrinsic spins by



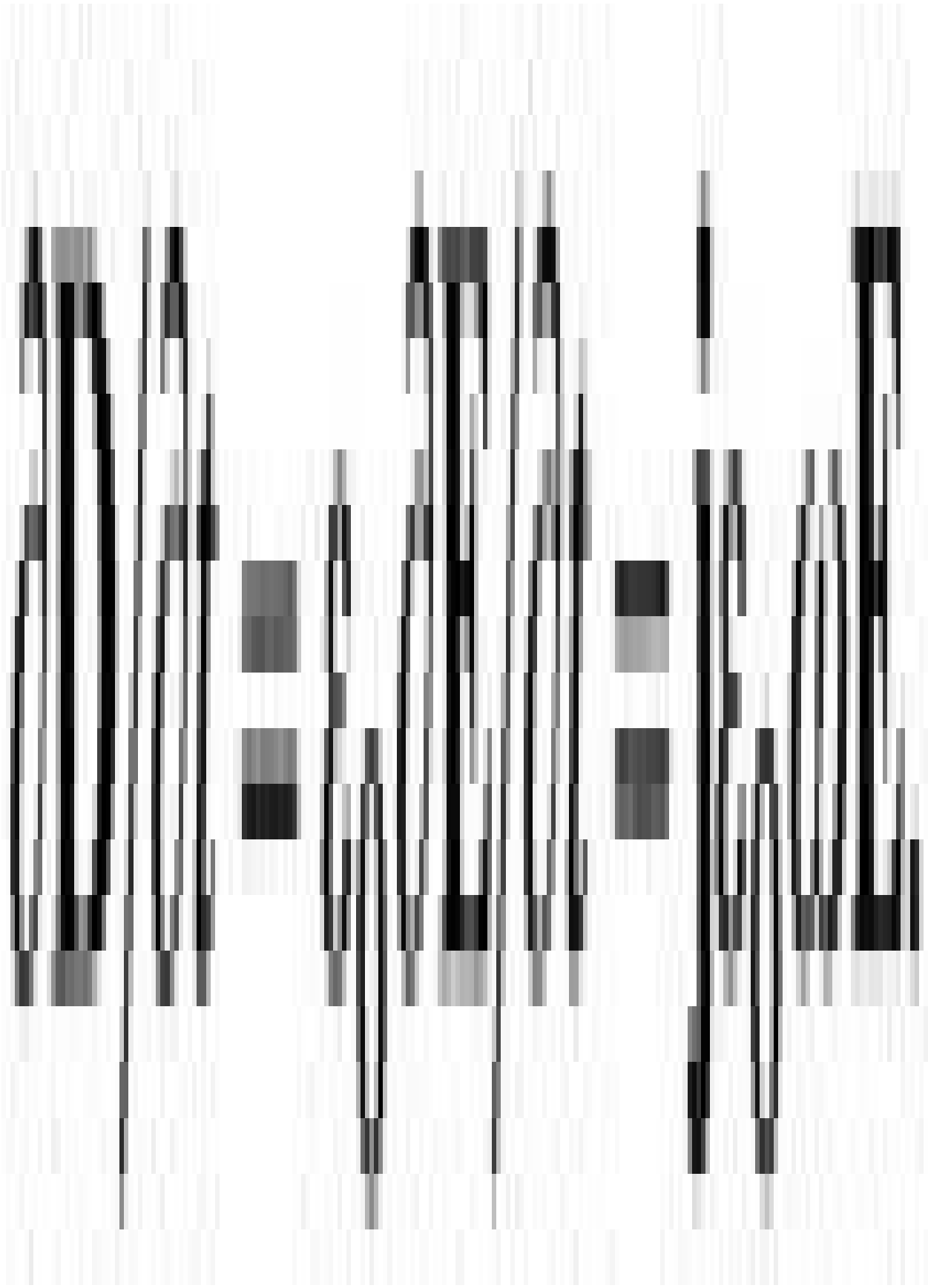
as discussed in Chapter 5. . is a constant for particles of the same type. Hence

where is a constant for particles of the same type.

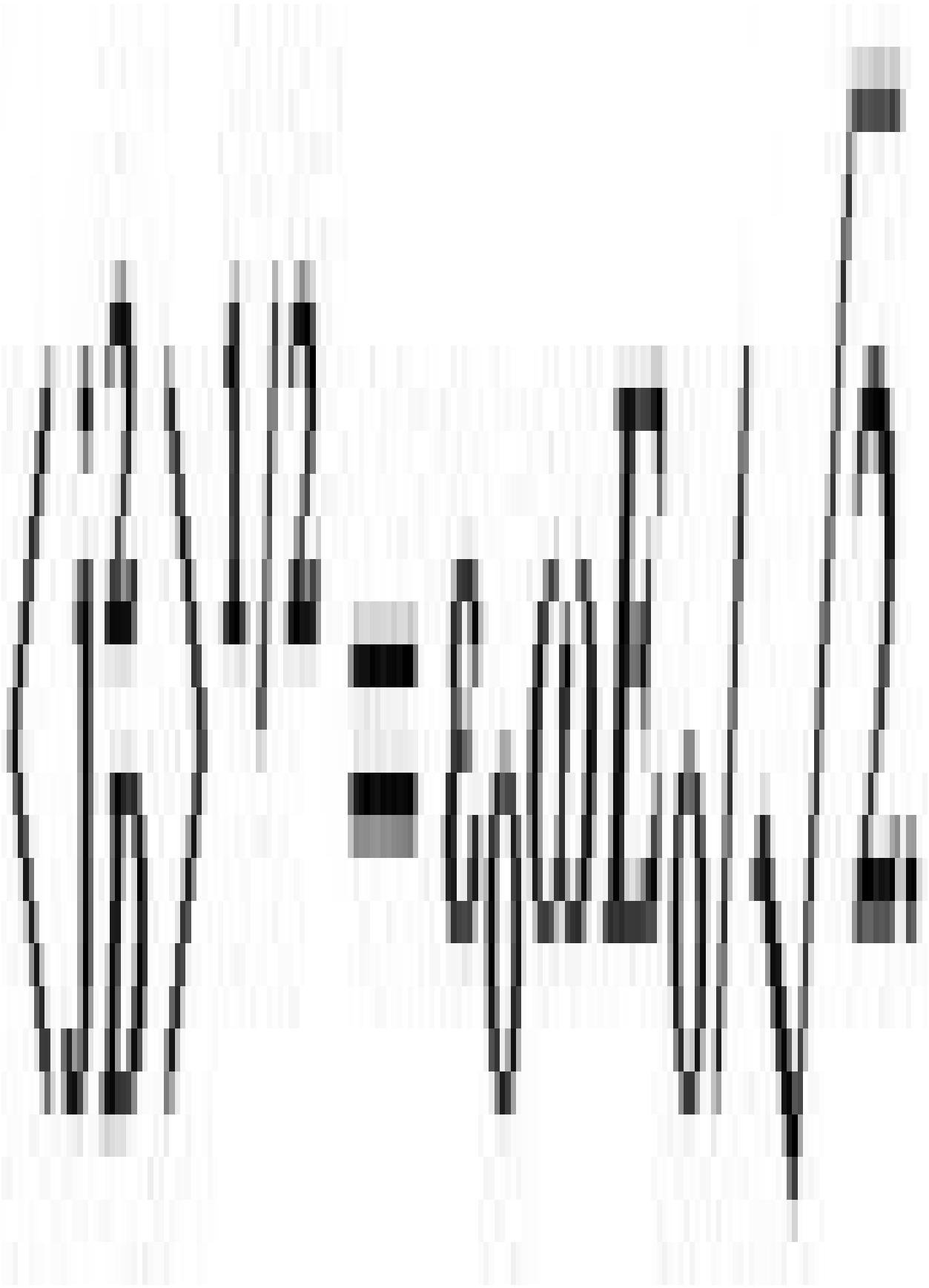
The electric dipole moment of an isolated atom or molecule usually vanishes, whereas the magnetic dipole moment is usually non-zero.

PROBLEMS 11

11.1 The displacement current density jD is equal to

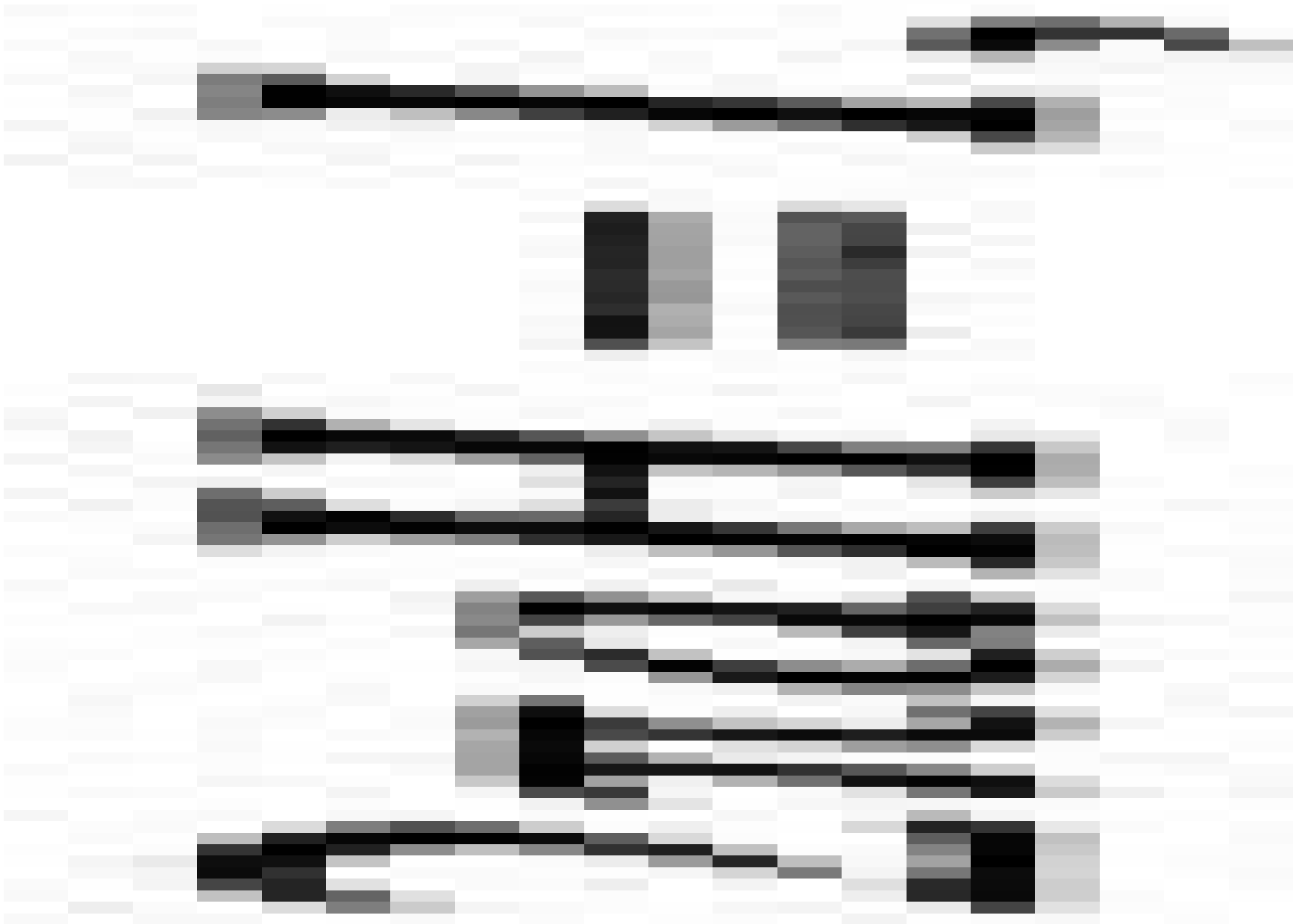


Hence

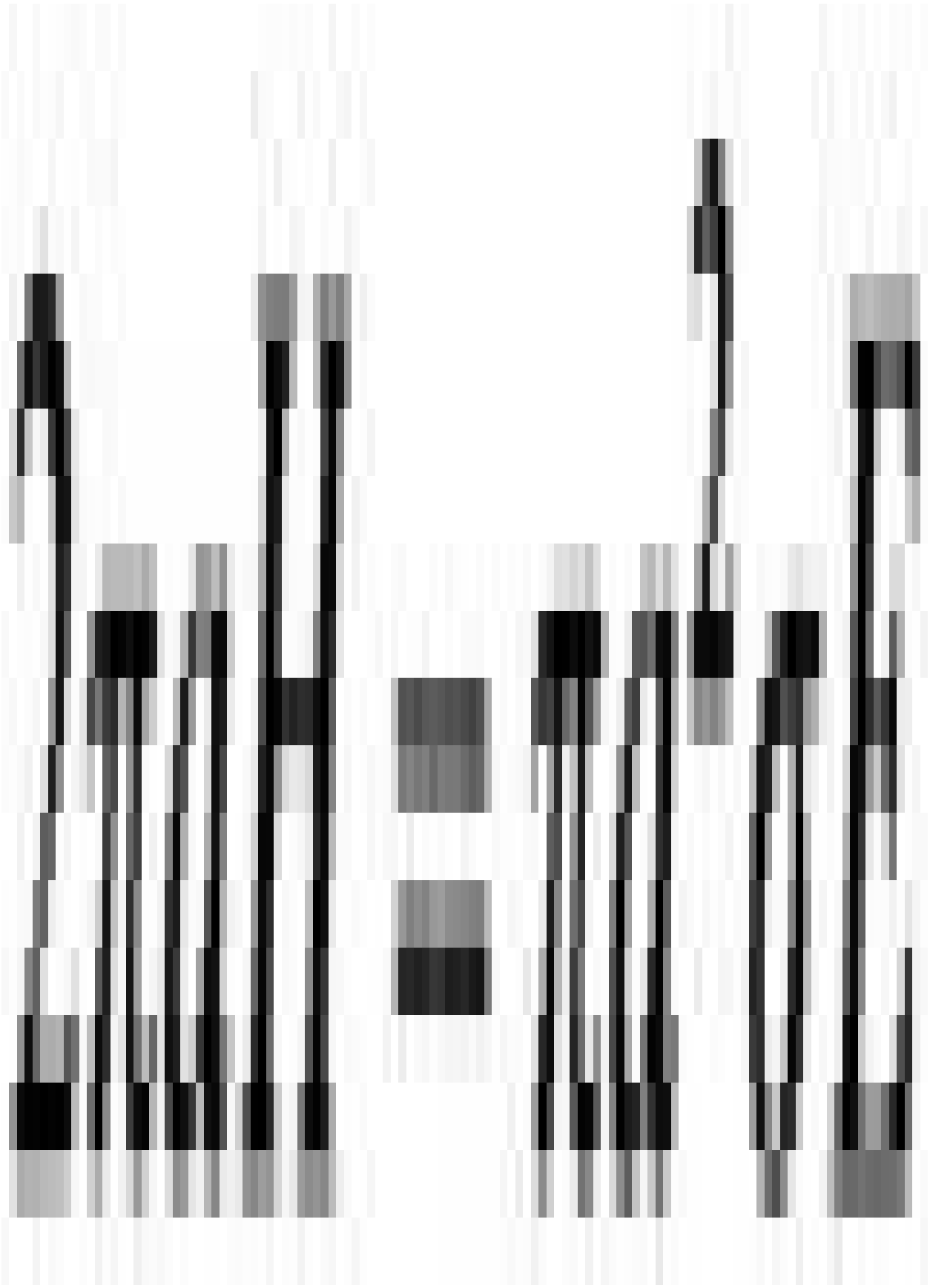


This gives E0 equal to 2.54×10^{-3} Vm^{-1}, and B0 equal to 8.47×10^{-12} T.

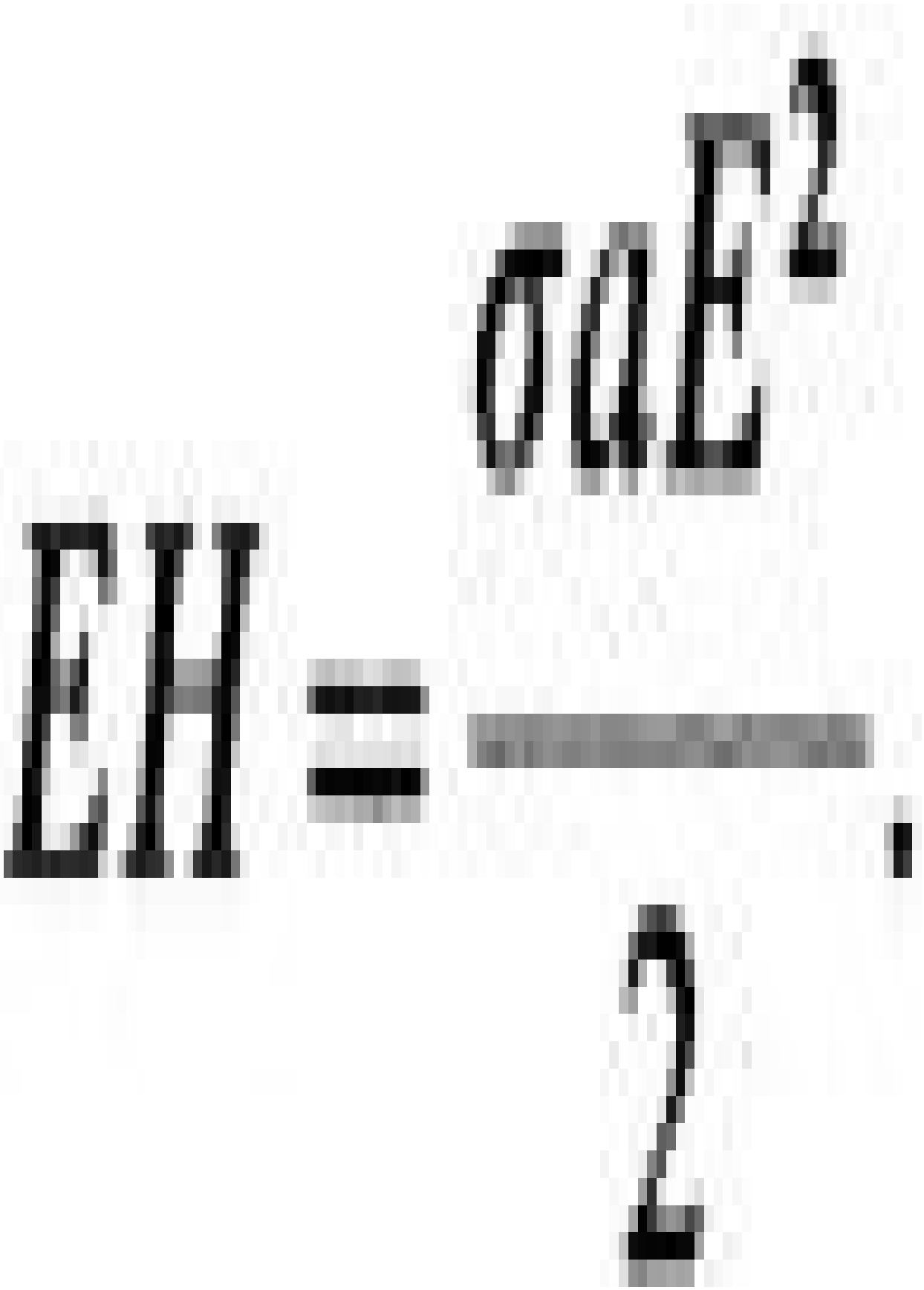
11.4 If the resistance of the wire is R and the capacitance is C the capacitor decays according to the expression $q = q_0 \exp(-t/RC)$. The field between the plates, assumed uniform, is thus given by $q = q_0 \exp(-t/RC)$, where $E_0 = V_0/l$, V_0 being the initial voltage across the plates and l their separation. By symmetry the magnetic field at the surface of the wire is tangential to the surface and has constant magnitude at a fixed time. By Ampère's law



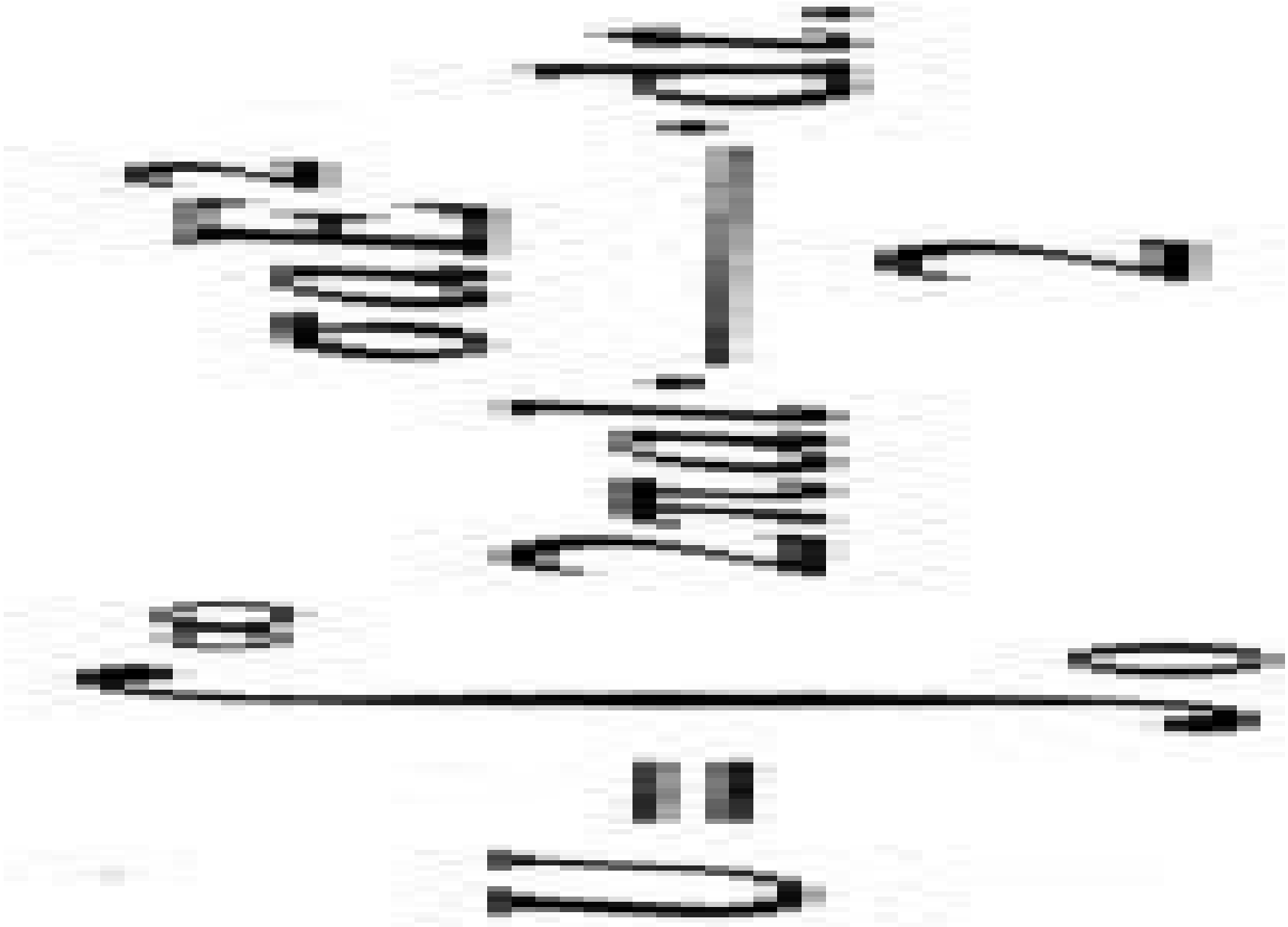
where a is the radius of the wire and I the current. We have neglected the displacement current through the wire; it is negligible if the conductivity $\sigma \gg \varepsilon_0\omega$, where ω may be taken to be of the order of $(RC)^{-1}$. The above equation may be written



and hence $H = \sigma\alpha E/2$. The Poynting vector $E \times H$ is constant over the surface of the wire and points inwards. It has magnitude



The total energy flowing into the wire is thus



This can be worked out to be equal to , the initial energy stored in the capacitor.

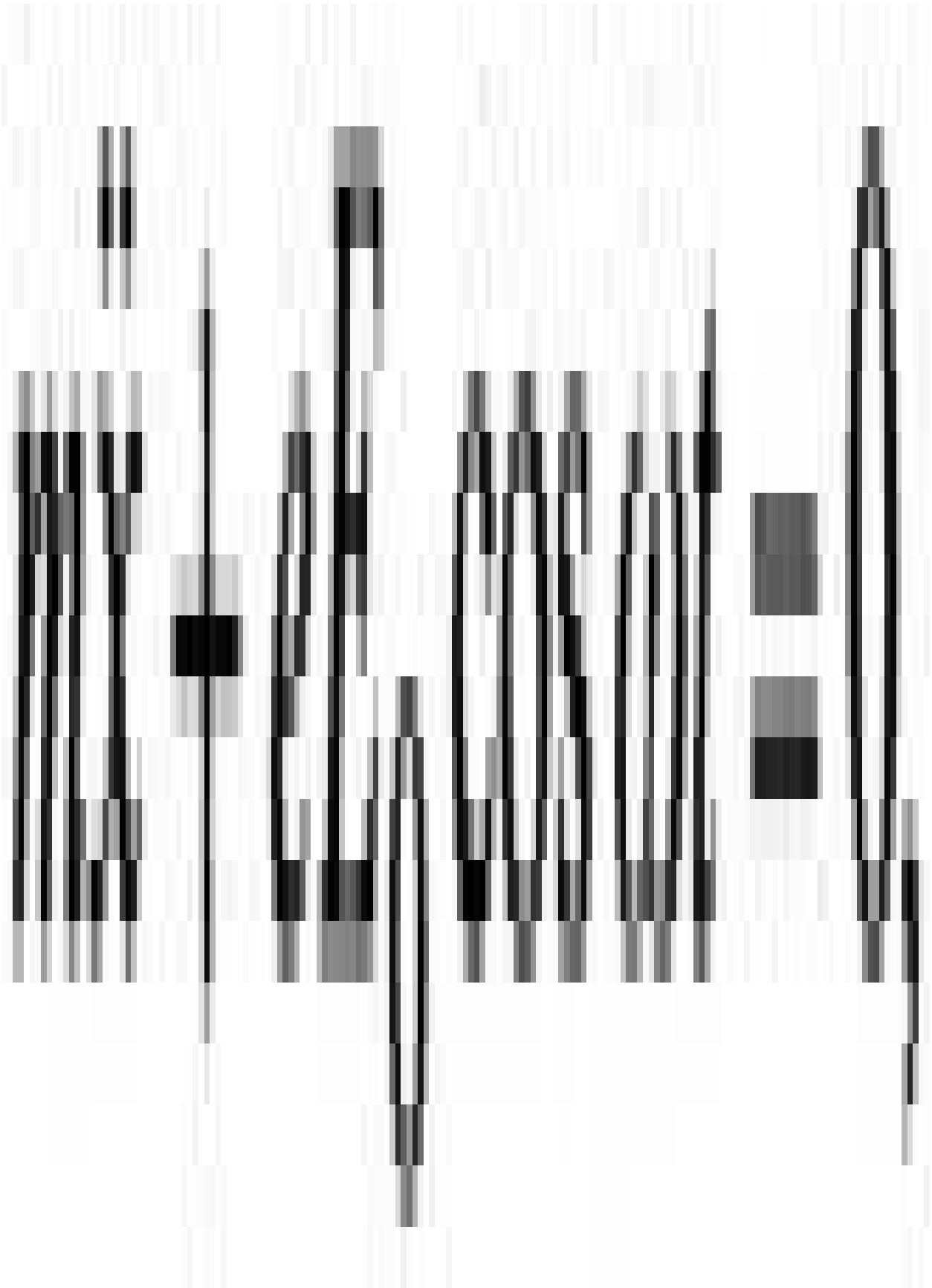
The use of the static field solution in time dependent problems can sometimes lead to difficulties. For example if we have a large parallel plate capacitor discharging through an external resistor the static field solution would suggest that curl E = 0 near the middle of the capacitor. From Faraday's law this implies that $\partial B/\partial t = 0$, and since the field B was initially zero it remains zero. Hence curl H would remain zero if our assumptions were correct. However this would imply that $\partial D/\partial t$ were zero, since curl $H = j + \partial D/\partial t$, and j is zero inside the capacitor. This conclusion contradicts our initial supposition. The way out of the puzzle of course is to realize that the static field solution is not always a sufficiently good approximation to the actual fields existing in time dependent situations.

11.5 The conductivity is 2.8×10^{-2} $(\Omega m)^{-1}$.

11.6 The resonant angular frequency ω is 3.16×10^{6} s^{-1}. Calculate the skin depth in copper at this frequency: the answer is about 10^{-4} m. Assuming that the current flows uniformly in an outer layer of the wire of thickness equal to the skin depth, the resistance R is estimated to be about 1Ω. If L is the self-inductance of the coil, Equation (7.26) gives the Q of the circuit to be $\omega L/R$, which is about 3000.

11.8 0.796 m.

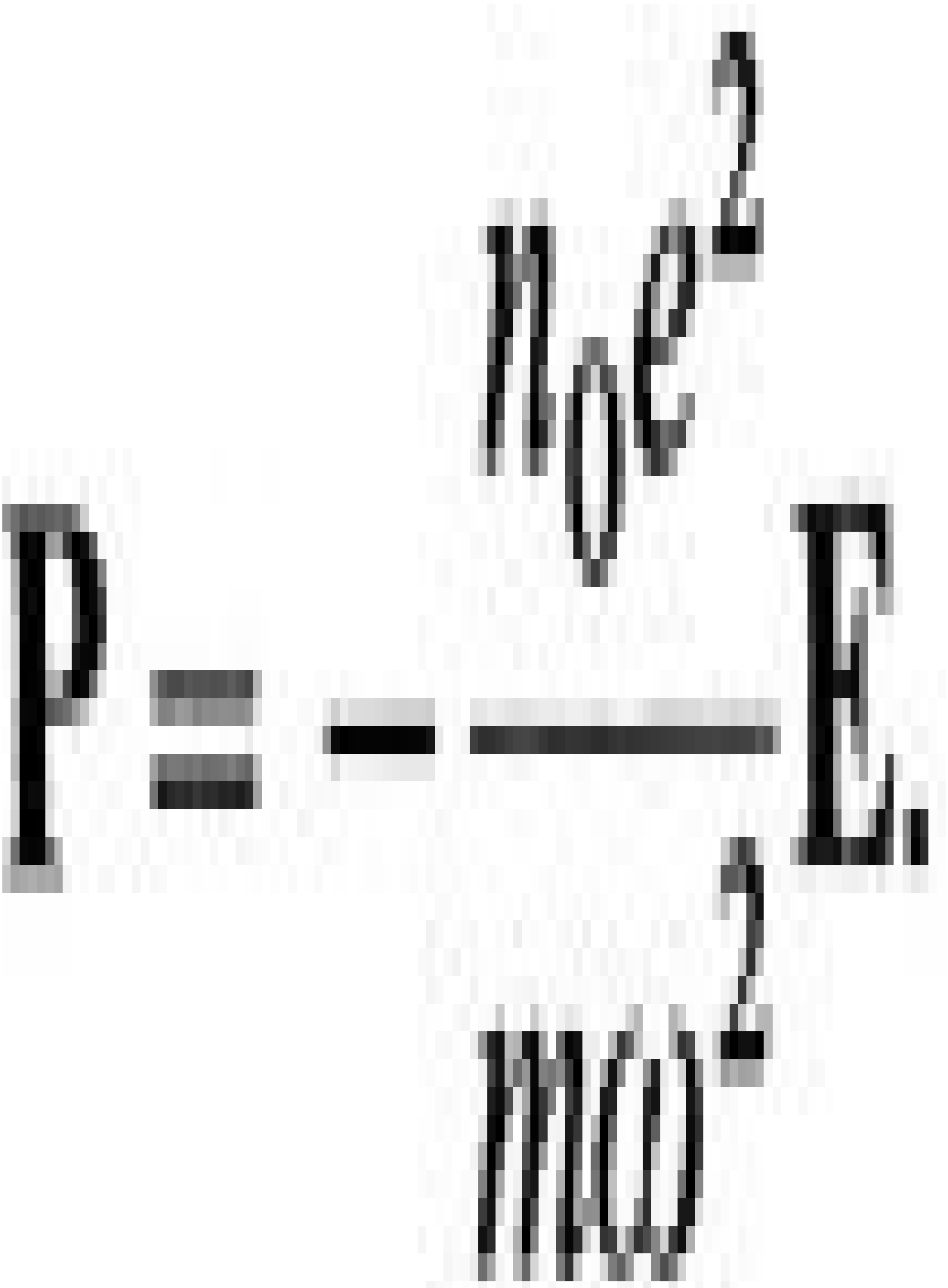
11.9 If the electric field is $E = exE0 \cos \omega t$ the equation of motion of an electron of mass m is



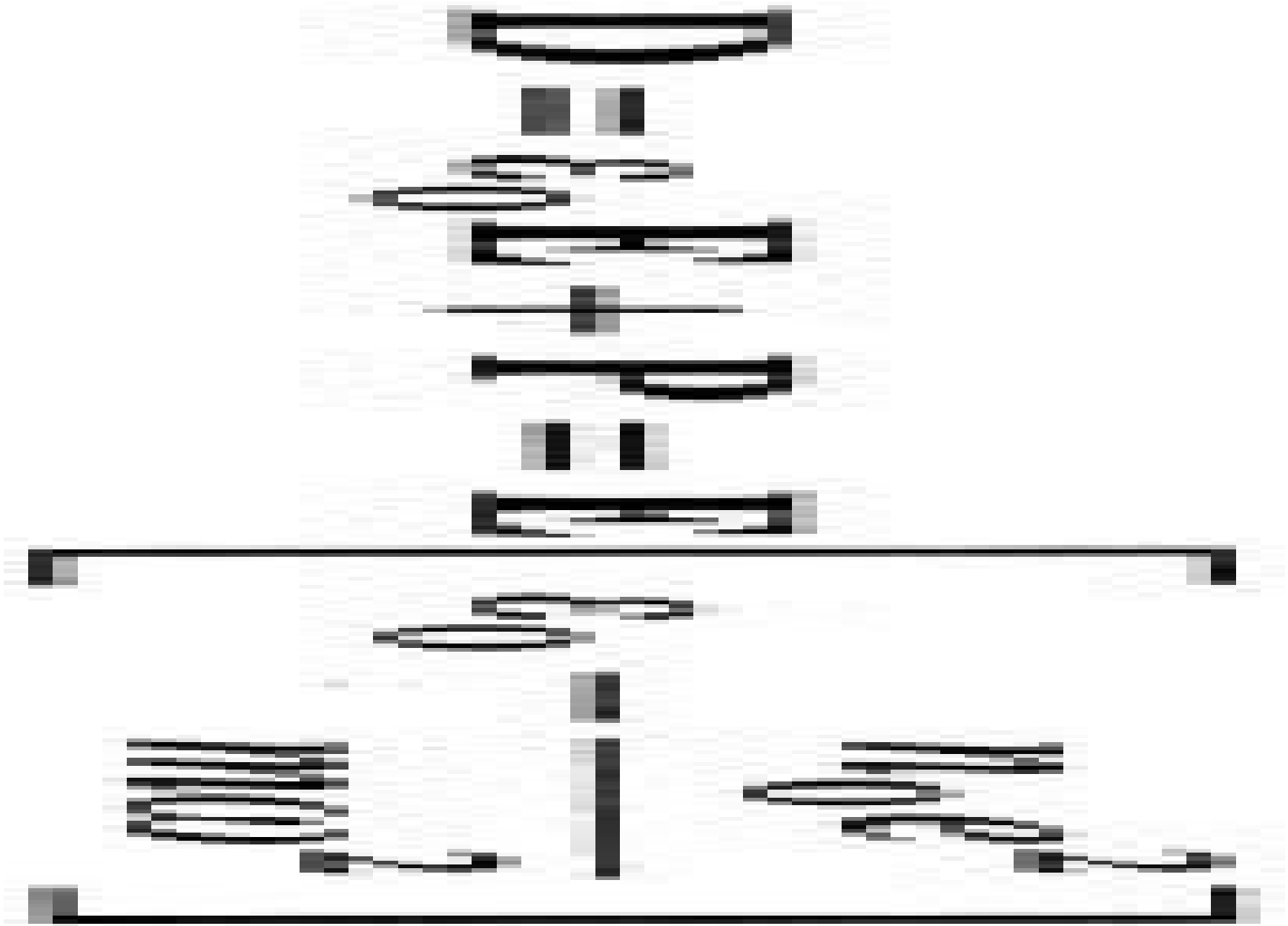
whence

$$x = -\frac{eE_0}{m\omega^2}\cos\omega t$$

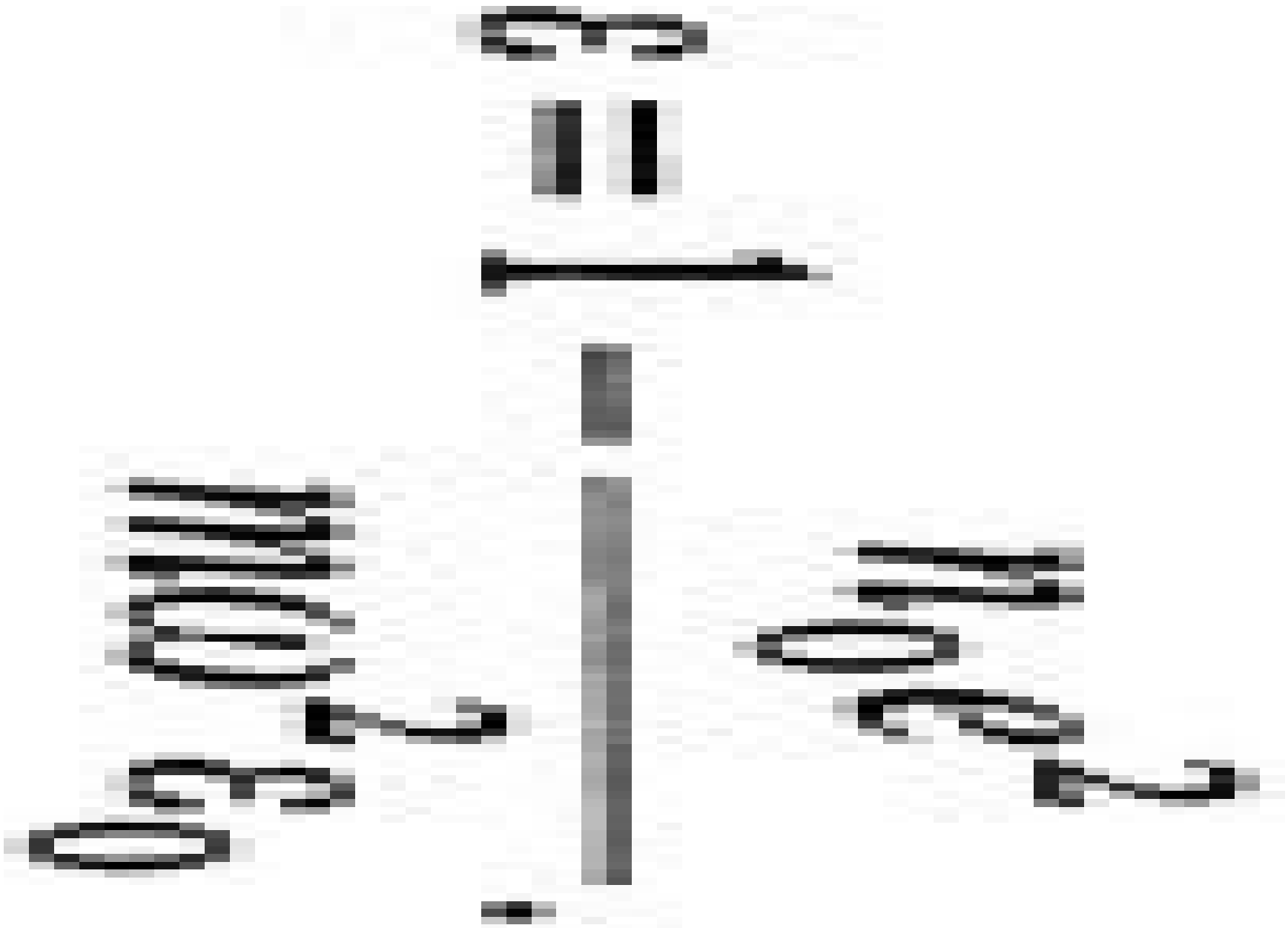
plus other terms which depend on the initial conditions. Since we are going to average over the n0 electrons in unit volume to obtain the polarization P, these other terms average to zero and may be ignored.



Hence



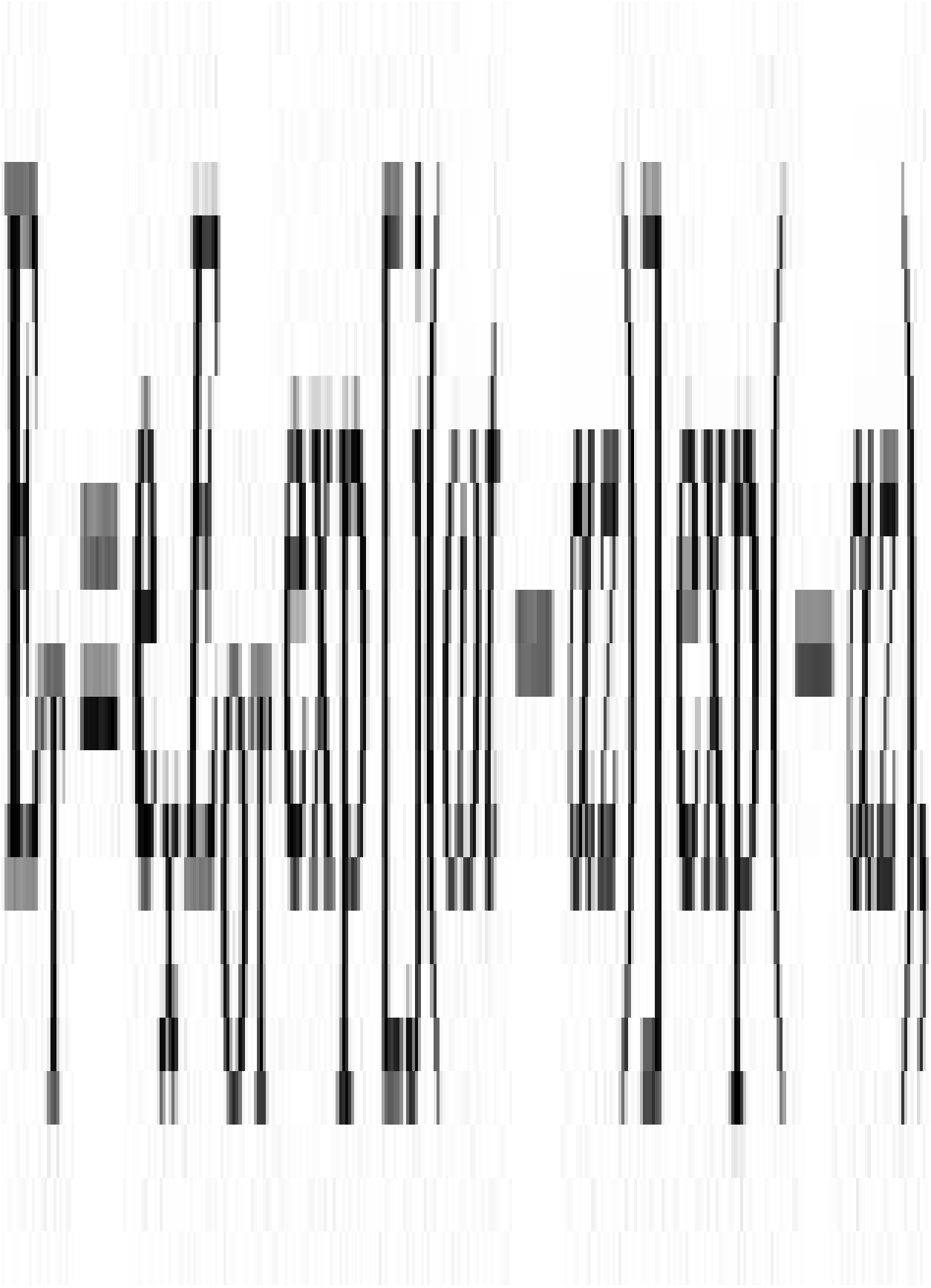
and



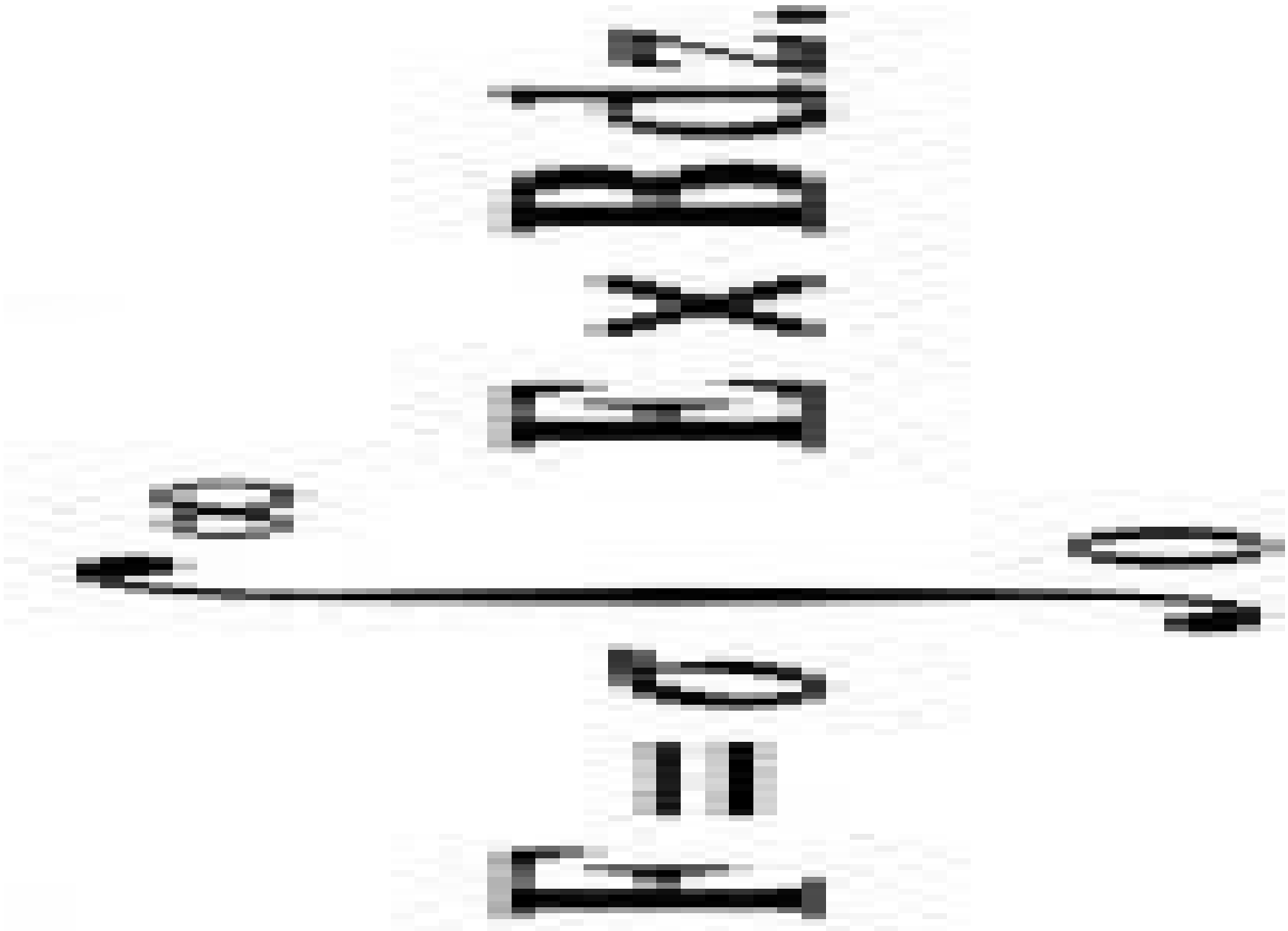
This gives the relative permittivity less than unity, and so the phase velocity of a wave in the ionosphere is greater than the speed of light in free space.

11.11 The water molecule is polar, and the permanent electric dipole moment contributes to the electric field as long as the molecule can rotate and keep aligned with the applied field. As soon as the frequency passes into the IR region the molecules can no longer rotate quickly enough to follow the field, because of their inertia. The relative permittivity thus begins to decrease. The nitrogen molecule is non-polar and its contribution to an electric field arises solely from the induced dipole moment. The frequencies at which such contributions begin to fall off depend on the electrons in the molecule and are in the visible or UV region of the electromagnetic spectrum.

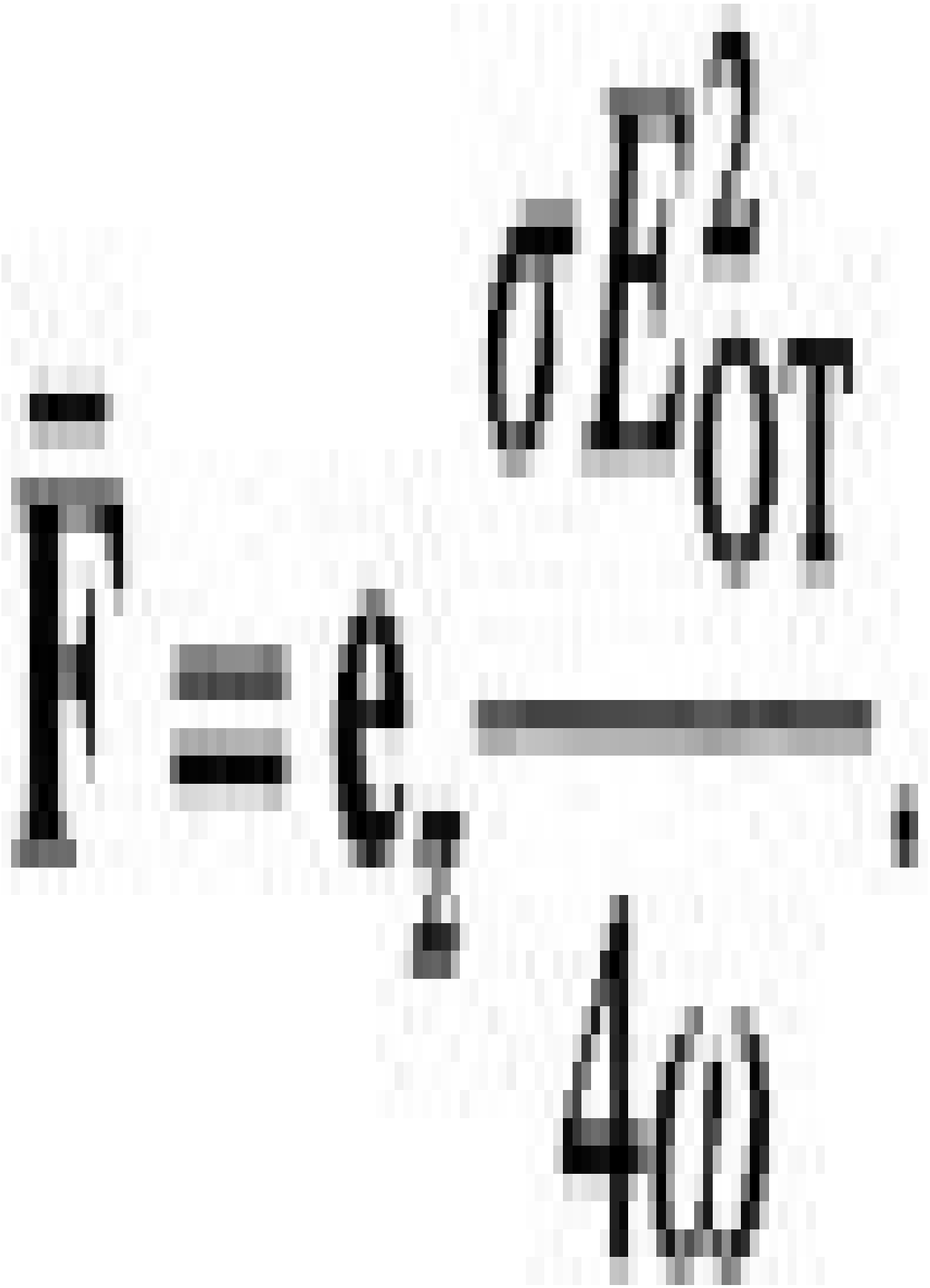
11.13 Let the boundary be at coordinate $z = 0$ and the electric wave in the metal be



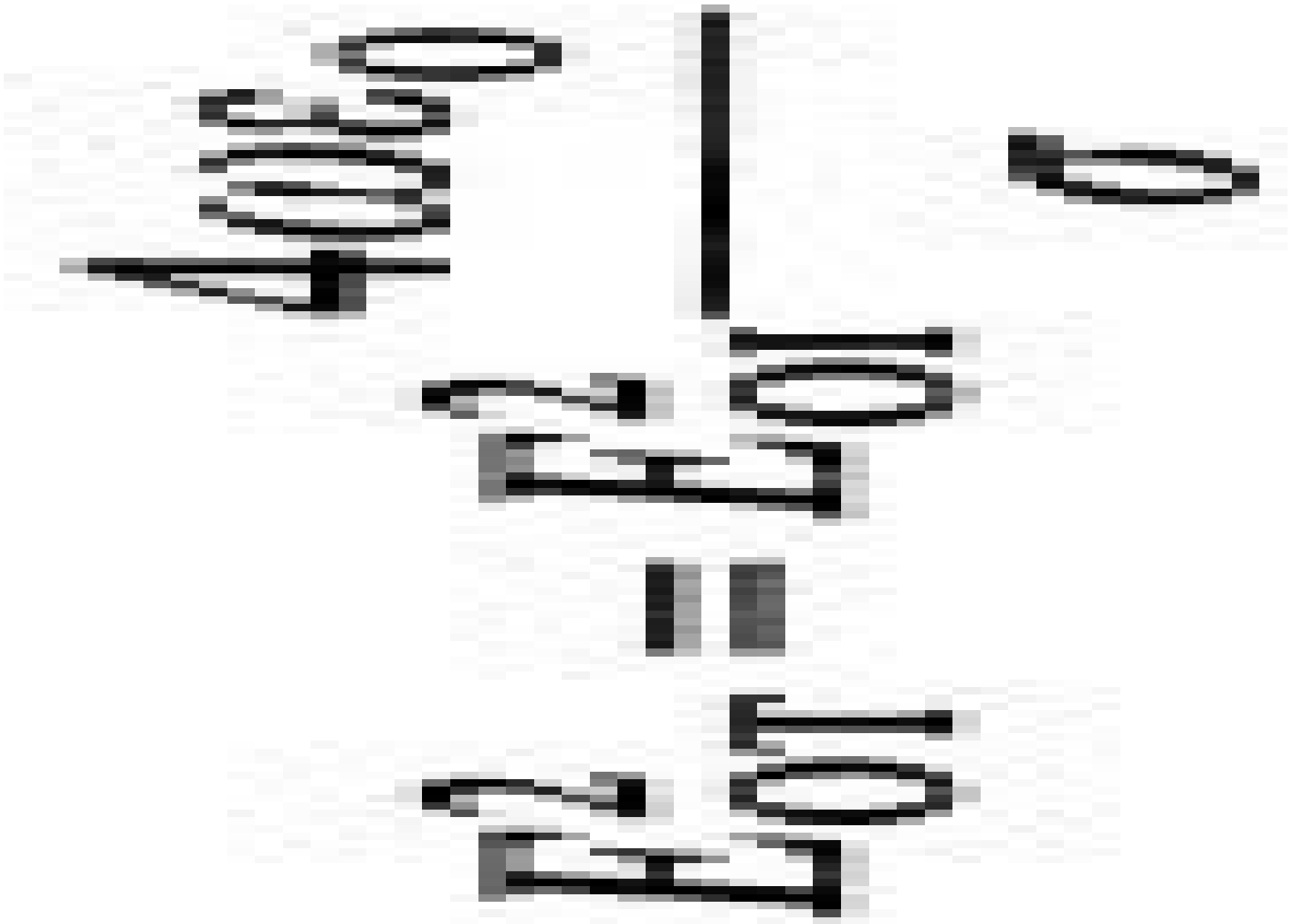
as in Equation (11.63). This electric wave produces a current density $j = \sigma E$ in the metal; $j = -Nev$, where N is the number of free electrons per unit volume and v is their mean drift velocity in the direction of the field E. The average force on an electron is thus $-ev \times B$ arising from the magnetic field in the wave. In a thickness dz of the metal at coordinate z the number of electrons per unit area is N dz and the force on them is thus $-Nev \times B\, dz$, which is equal to $\sigma E \times B\, dz$. The total force on unit area of the metal is



The fields E and B can be obtained from the expressions given in section 11.6.3, and the average force calculated. The answer obtained is



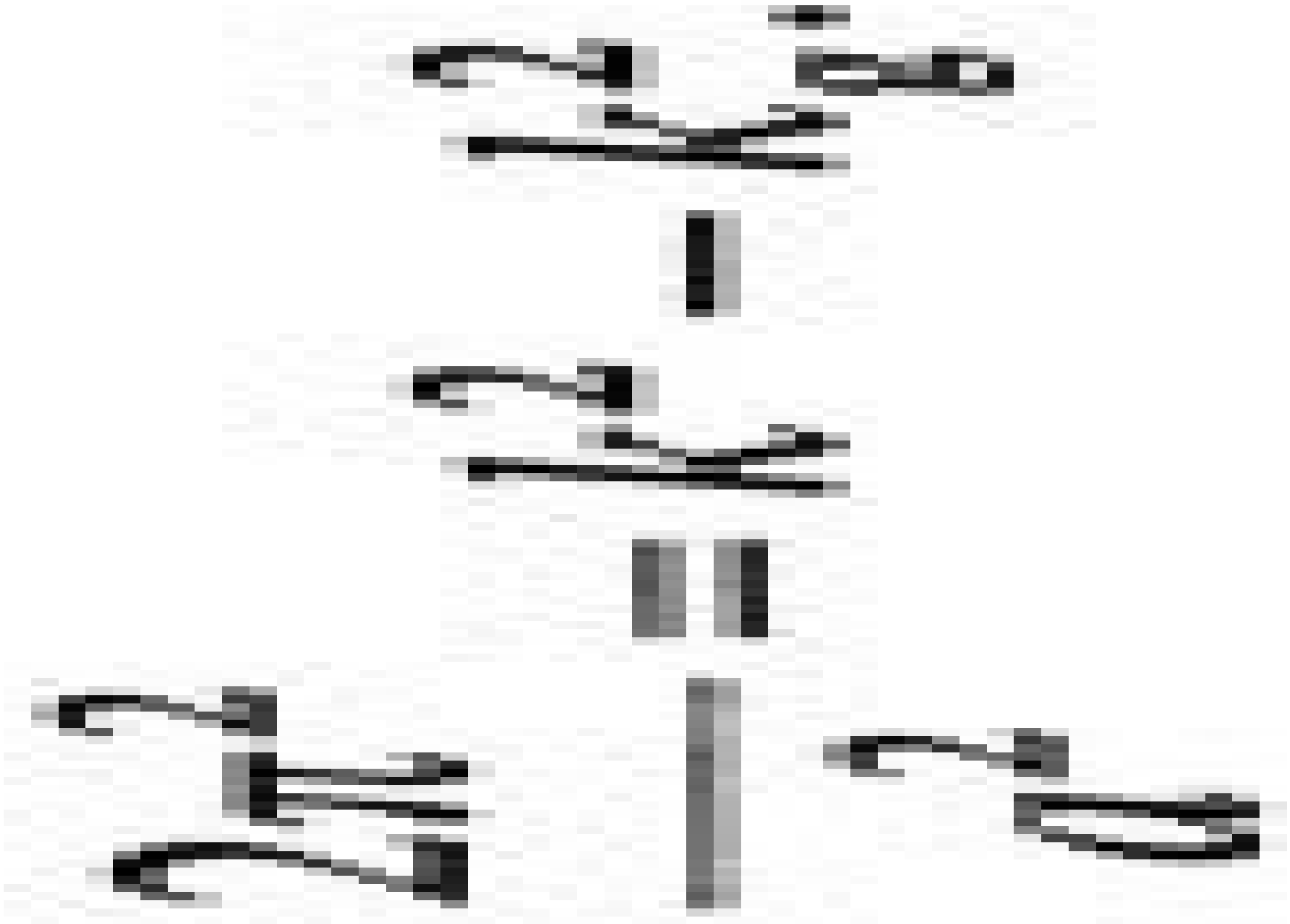
We now have to return to the considerations of section 11.6.3 to determine the relationship between EOT and EOI, the amplitude of the electric field in the wave incident on the metal. As the conductivity σ tends to infinity we find



and hence , the same result as Equation (11.74).

PROBLEMS 12

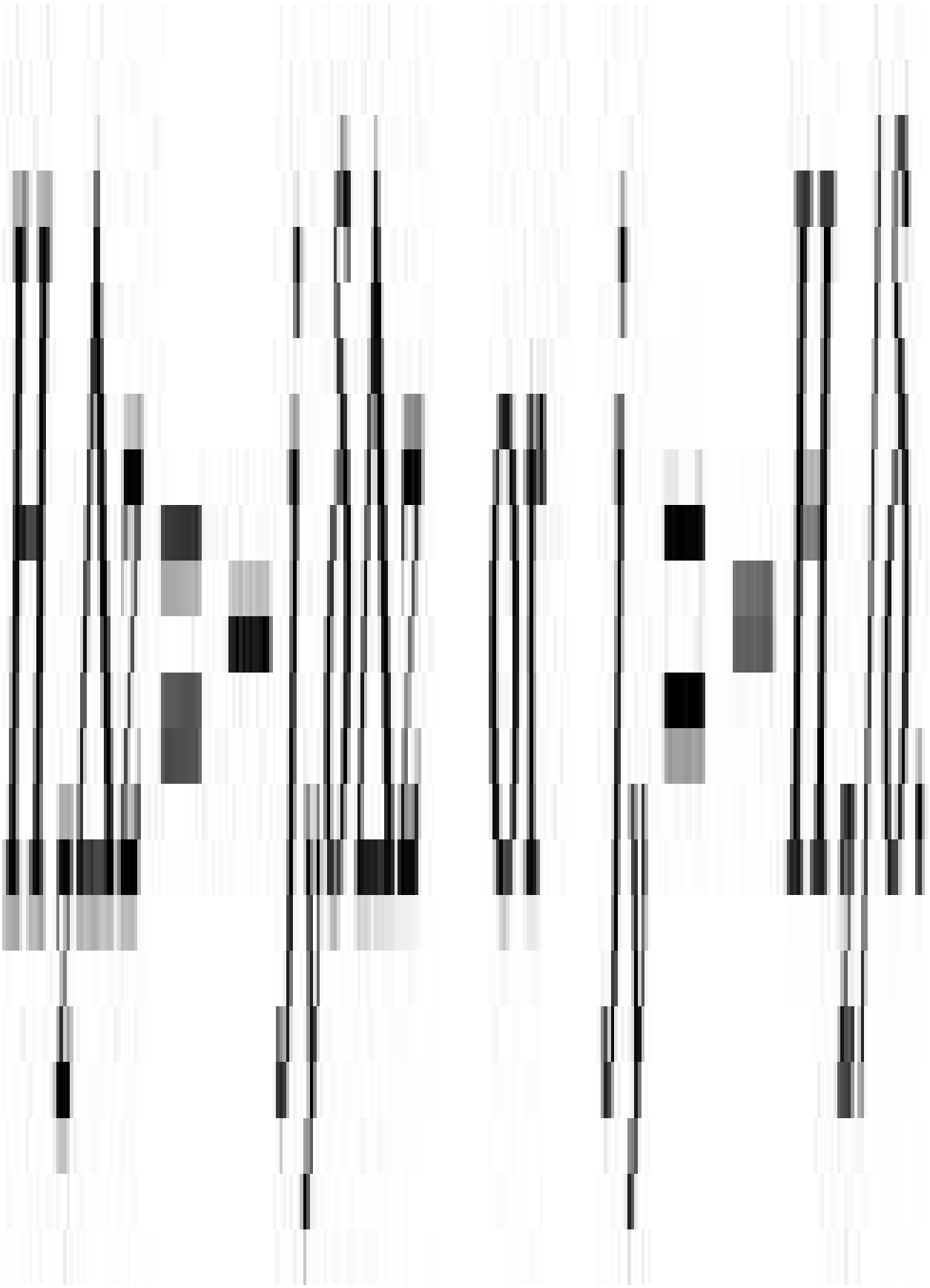
12.2 If a is the length of a side of the waveguide, from Equation (12.18) we see that no wave at all will be propagated without attenuation if $\pi/a > k$. Hence the minimum value of a is given by $\pi/k = \lambda/2$. The TE11 wave has the next lowest cut-off frequency and obeys the equation



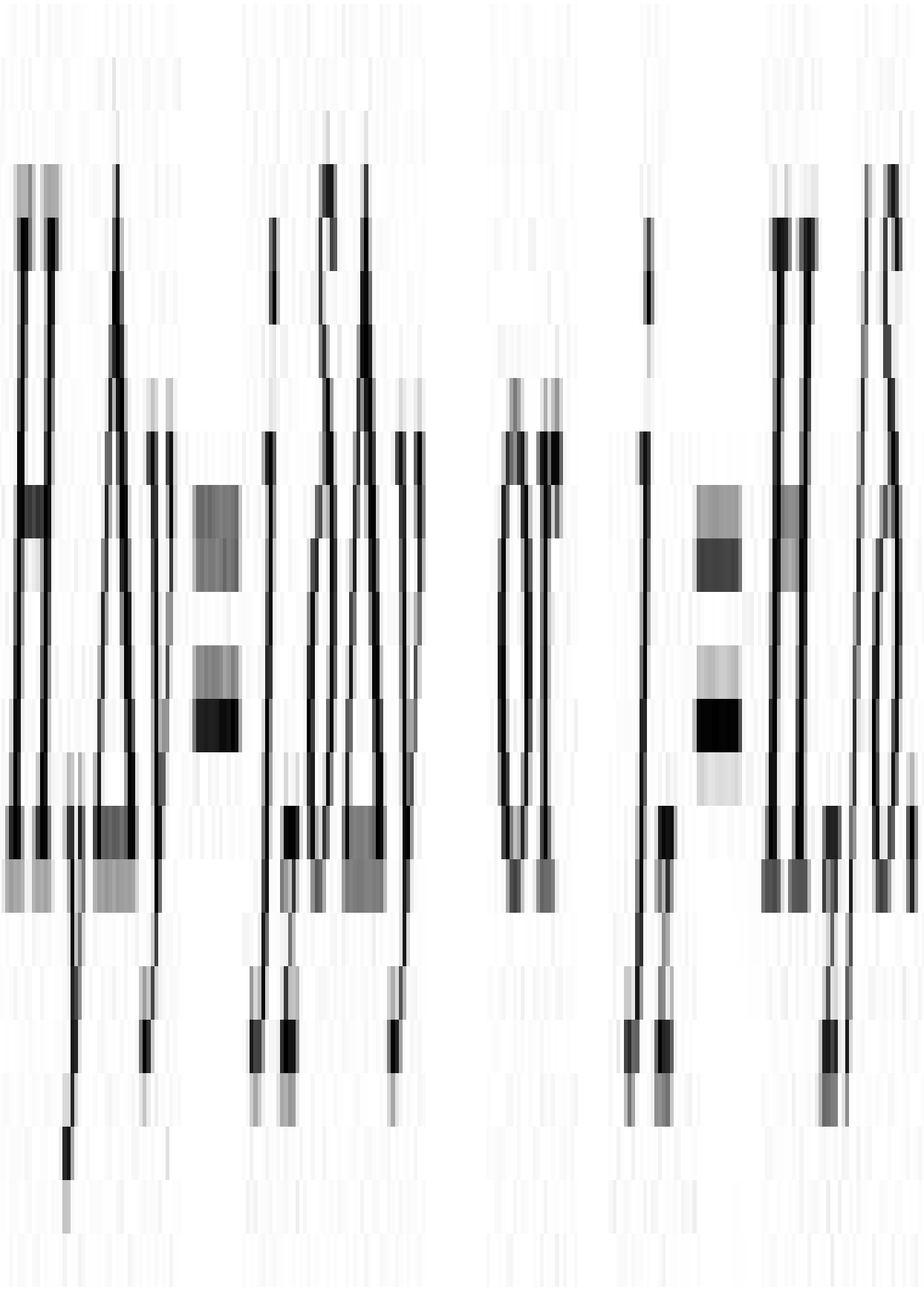
Hence if $2\pi^2/a^2 > k^2$ the TE11 wave will not propagate, i.e. if no wave other than the TE01(TE10) will propagate. Hence the maximum value of a is .

12.3 A similar calculation was done on p. 421. The answer is 0.79 MW.

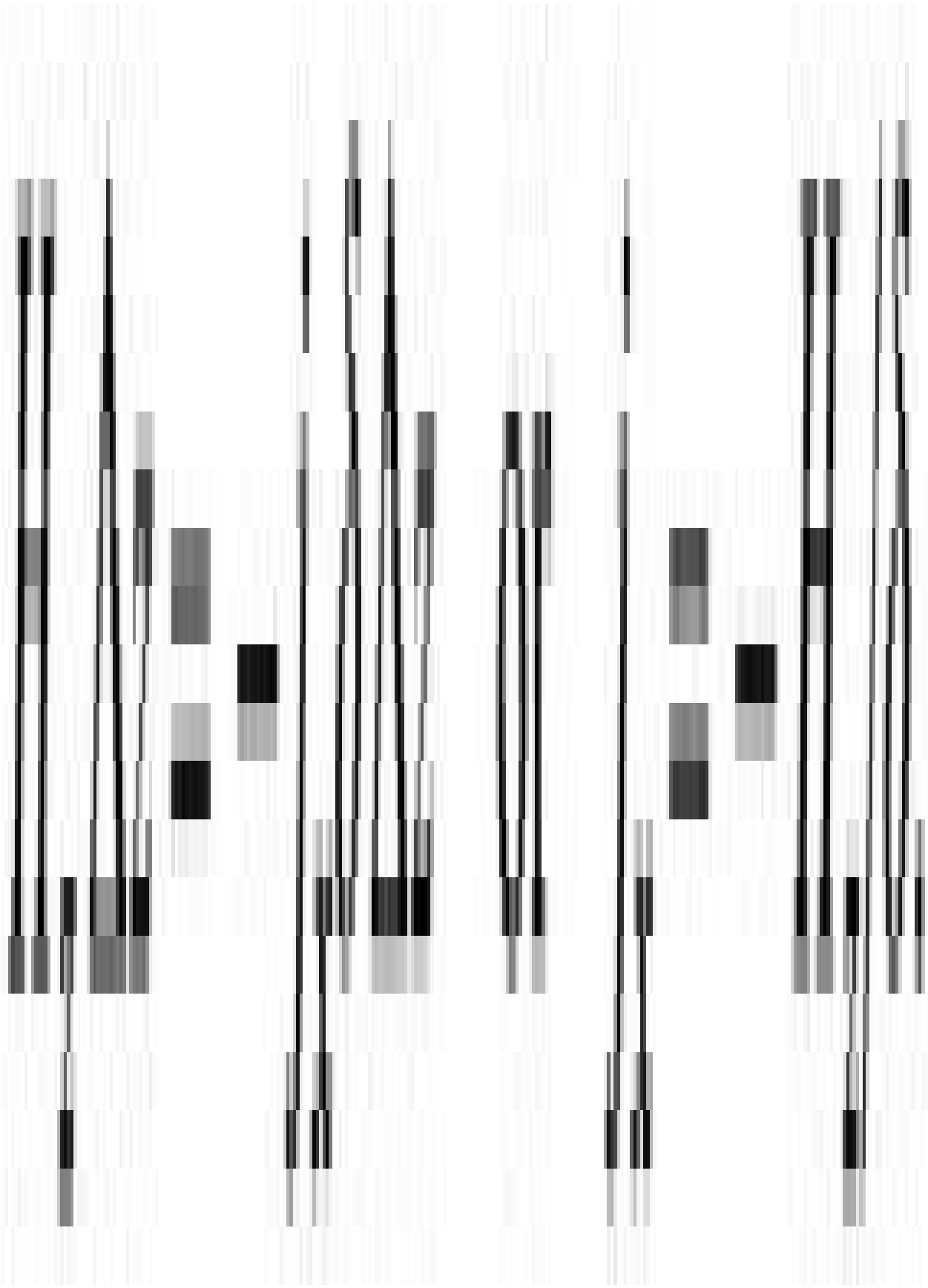
12.6 We can make an estimate of the losses by assuming that the currents in the walls flow only in a thin layer of thickness equal to the skin depth δ, and that the current density is uniform within this layer. Applying Ampère's integral law (5.29) over a loop of small length Δz and height δ lying along the z-direction in the top or bottom walls, we have



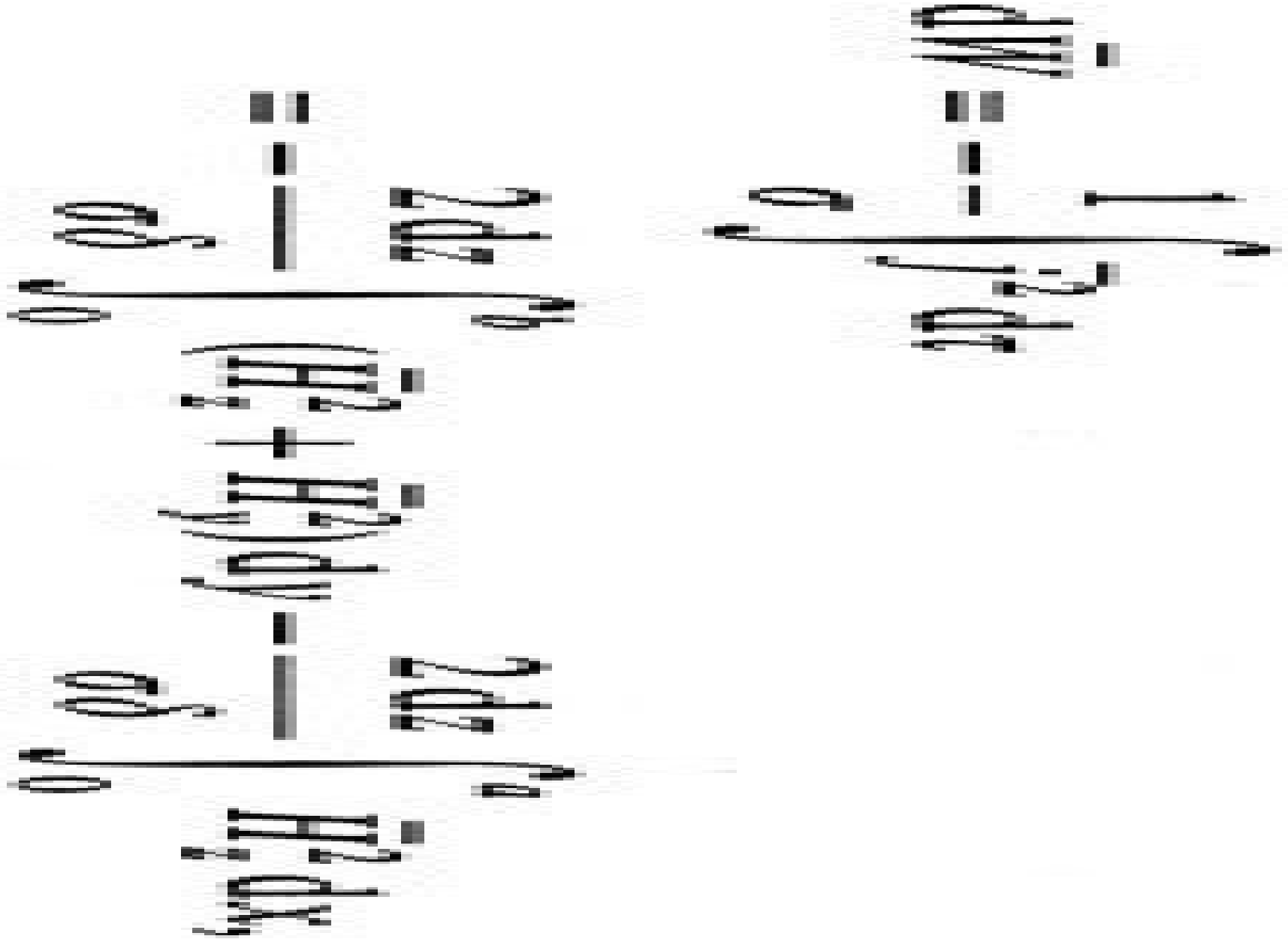
Applying the law to a loop of length Δy, height δ lying along the y-direction in the top or bottom wall,



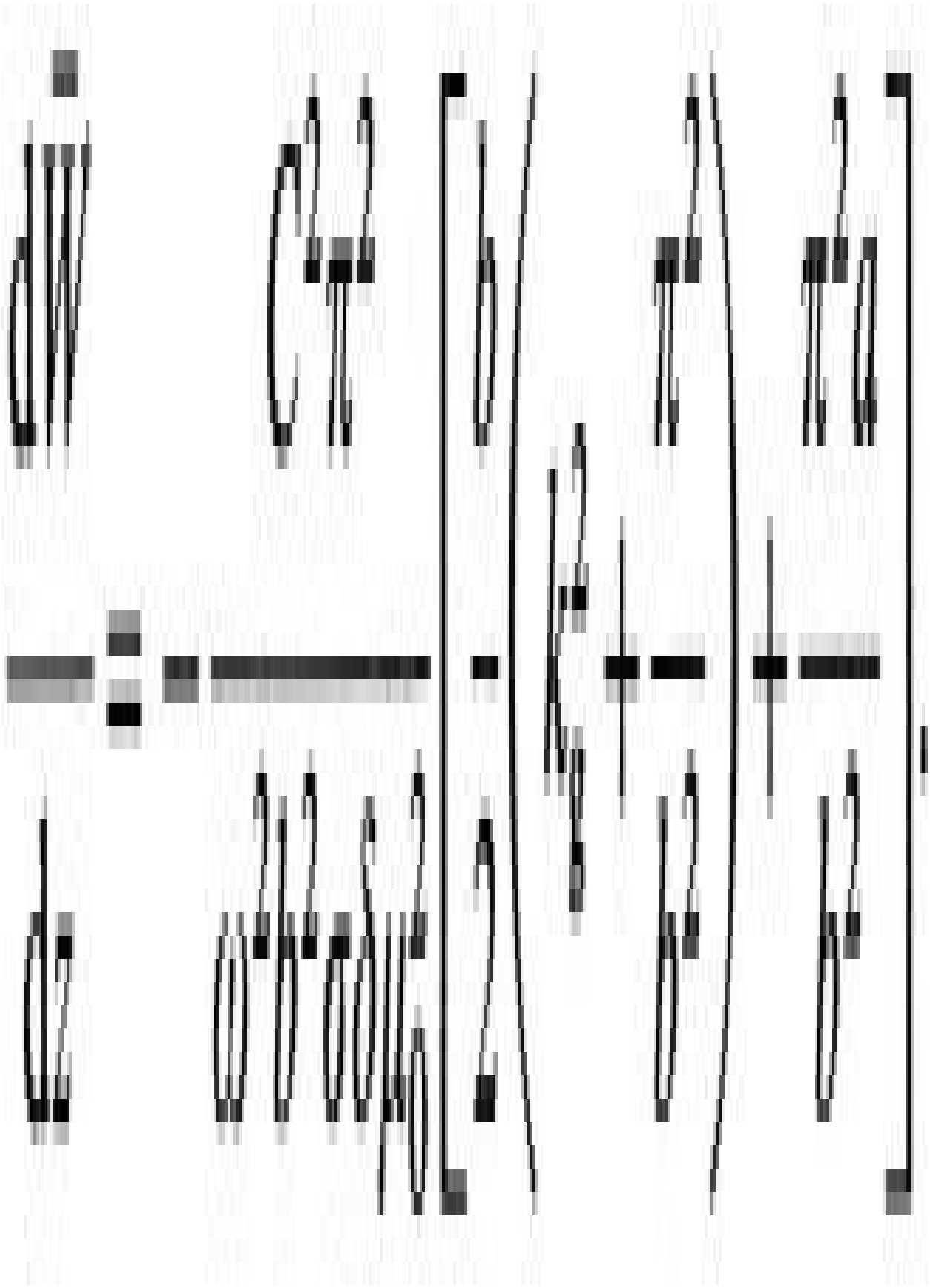
Applying the law to a loop of length Δz height δ lying along the z-direction in either of the side walls at $y = 0$ or $y = b$,



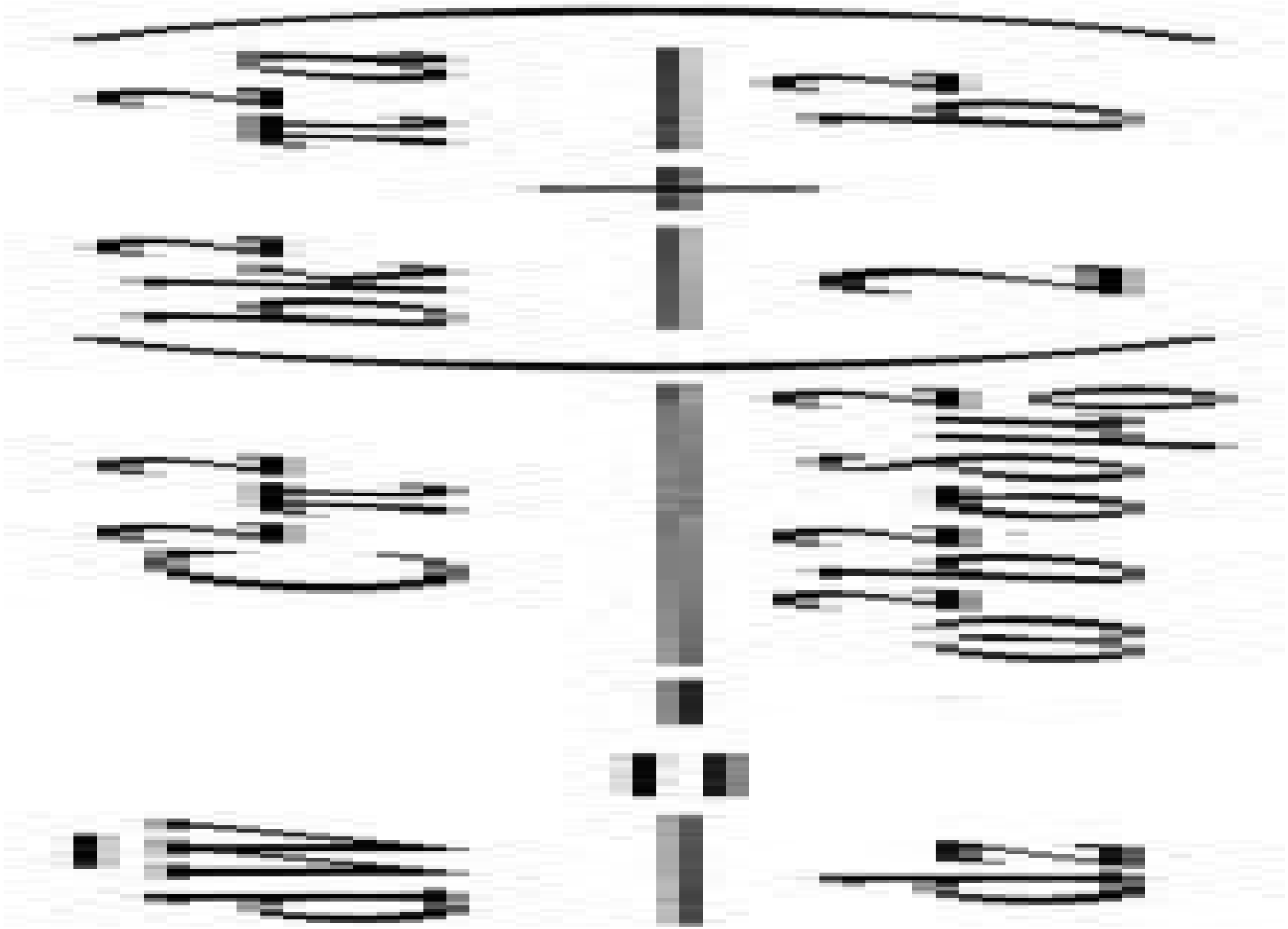
The average power loss d in the strip of thickness δ length dz around the guide is



where the first term comes from the top and bottom walls, and the second term comes from the side walls. Integrating, and using Equations (12.27) and (12.28) to obtain Hy and Hz, gives



From Equation (12.18)

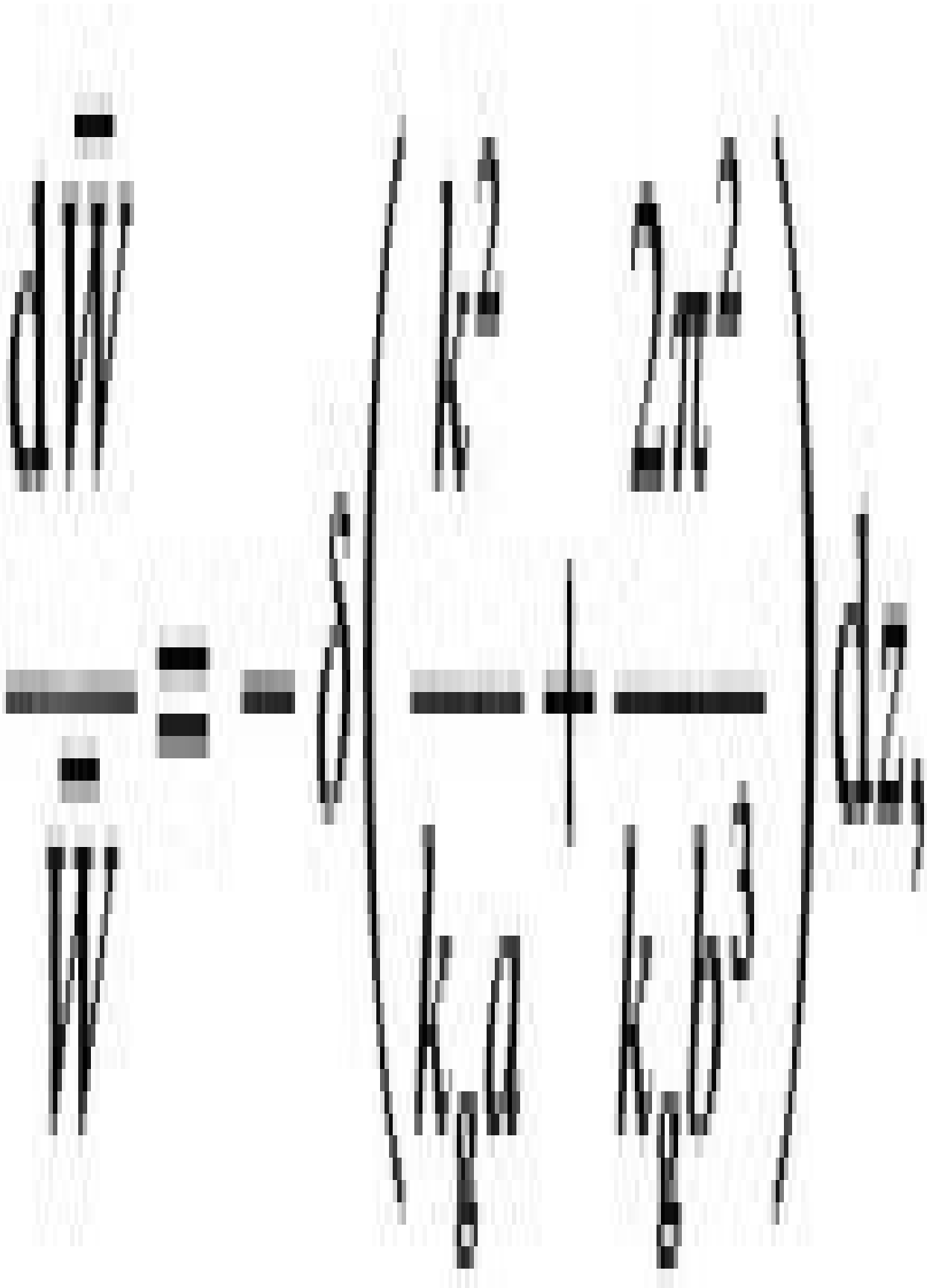


The average power flowing down the guide is

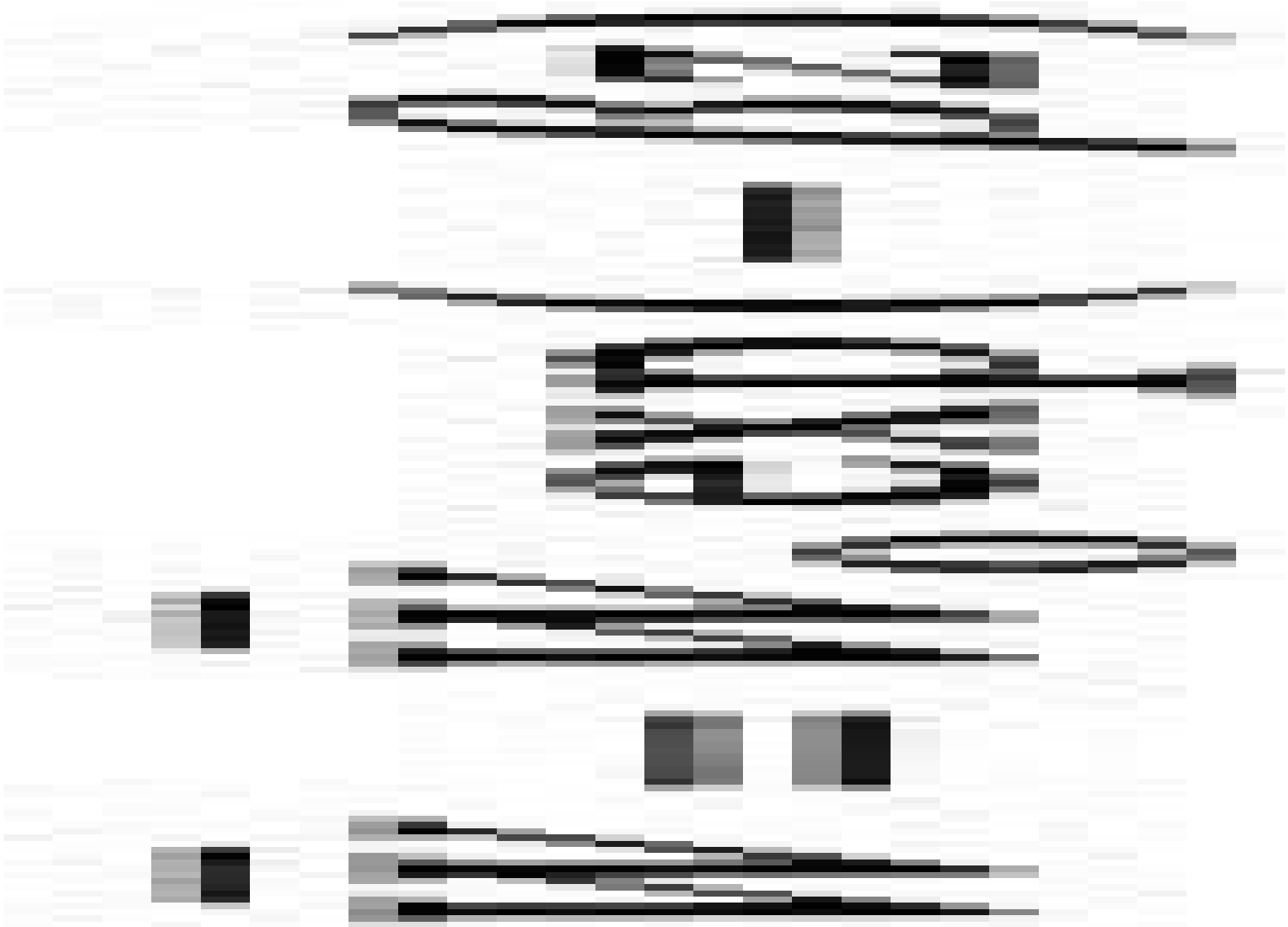
$$\bar{W} = \int_0^a \int_0^b N_g \, dx \, dy$$

$$= \frac{1}{4} \frac{C^2 \pi^2 k_g}{\omega b^2 \mu_0} ab.$$

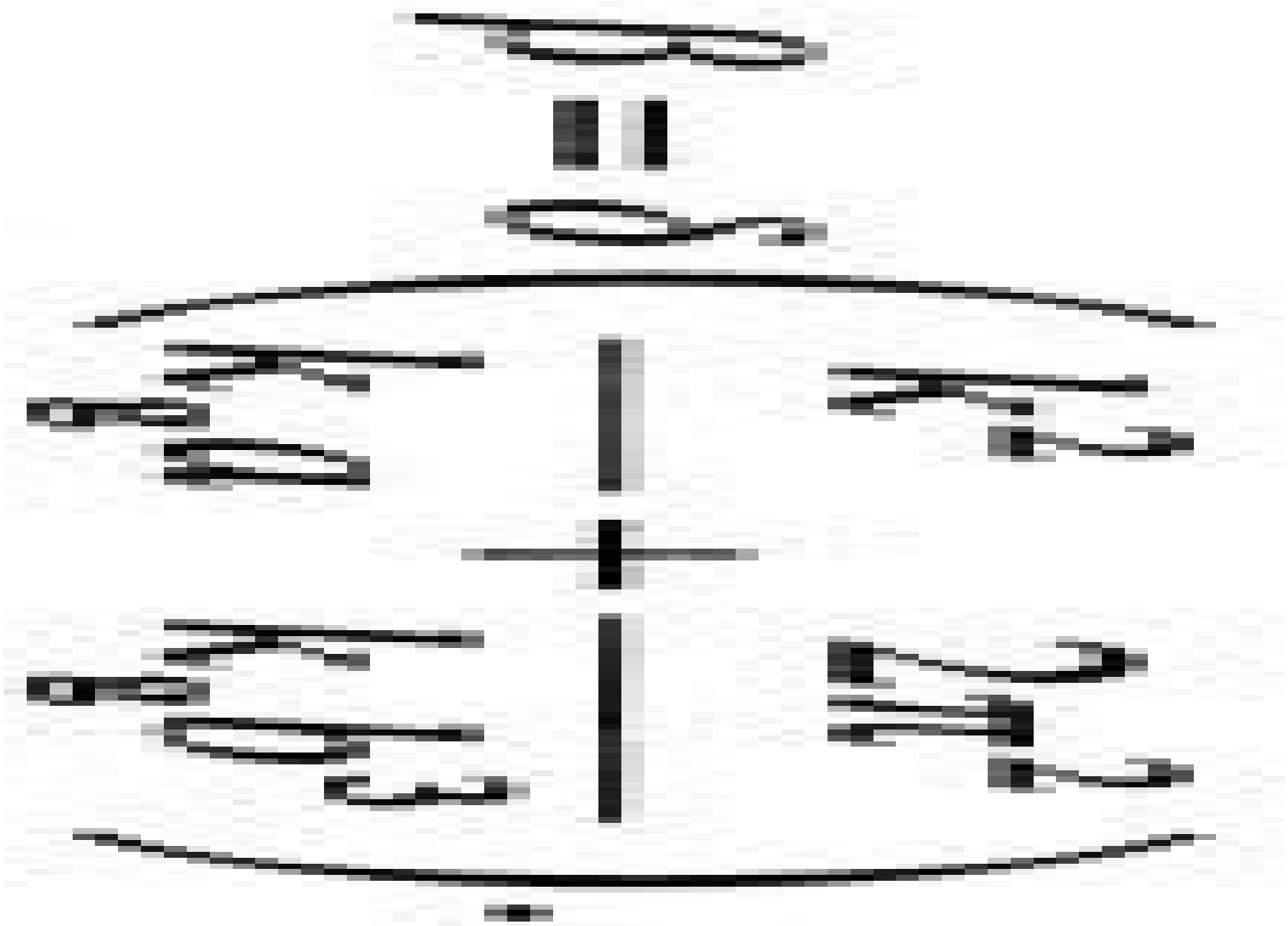
Hence



where we have used Equation (11.45) for the skin depth δ. The above equation leads to an attenuation constant β in the equation



given



Substitution of the appropriate values for 10^{10} Hz in a brass guide with a = 1.0 cm, b = 2.0 cm gives β equal to 0.06 m−1, i.e. the power falls to one half over a distance of about 0.7/β ~ 11m.

12.7 There are five resonances each of the TEmnl and TMmnl modes.

12.10 The angular frequency ω is equal to

PROBLEMS 13

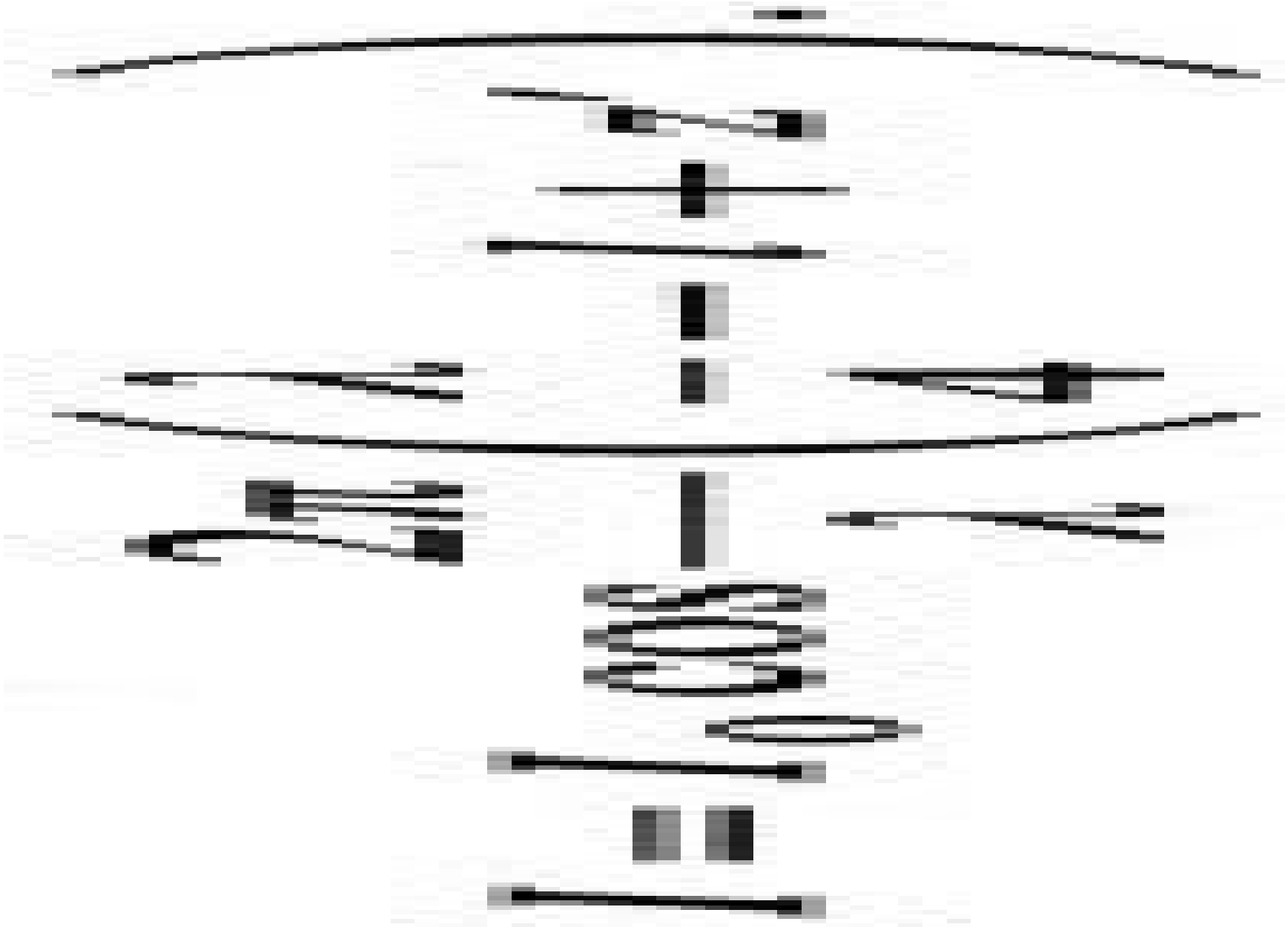
13.1 Answer is about 400 W.

The frequency chosen has to be above the international long-range communications band−the 10m band. There are only narrow frequency ranges allotted for taxis, police etc. A frequency of 10^8 Hz is a suitable frequency from the point of view of cost and simplicity of electronics, and size of antennas. The wavelength of 3 m also ensures that there is sufficient diffraction of the waves around buildings to give signals everywhere.

13.2 See section 9.4.

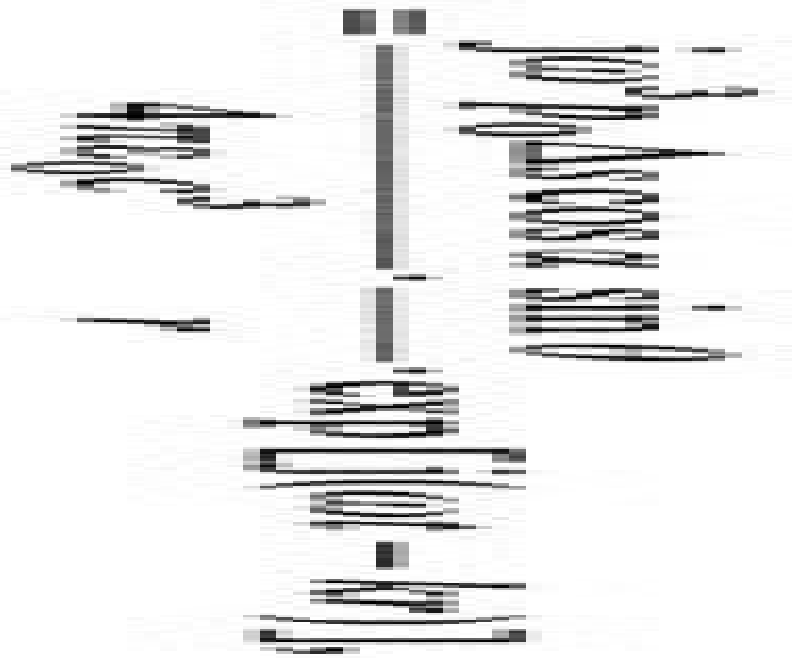
13.3 This problem is the four slit problem in optics where four narrow slits radiate in phase. The pattern is thus like that of a four line diffraction grating.

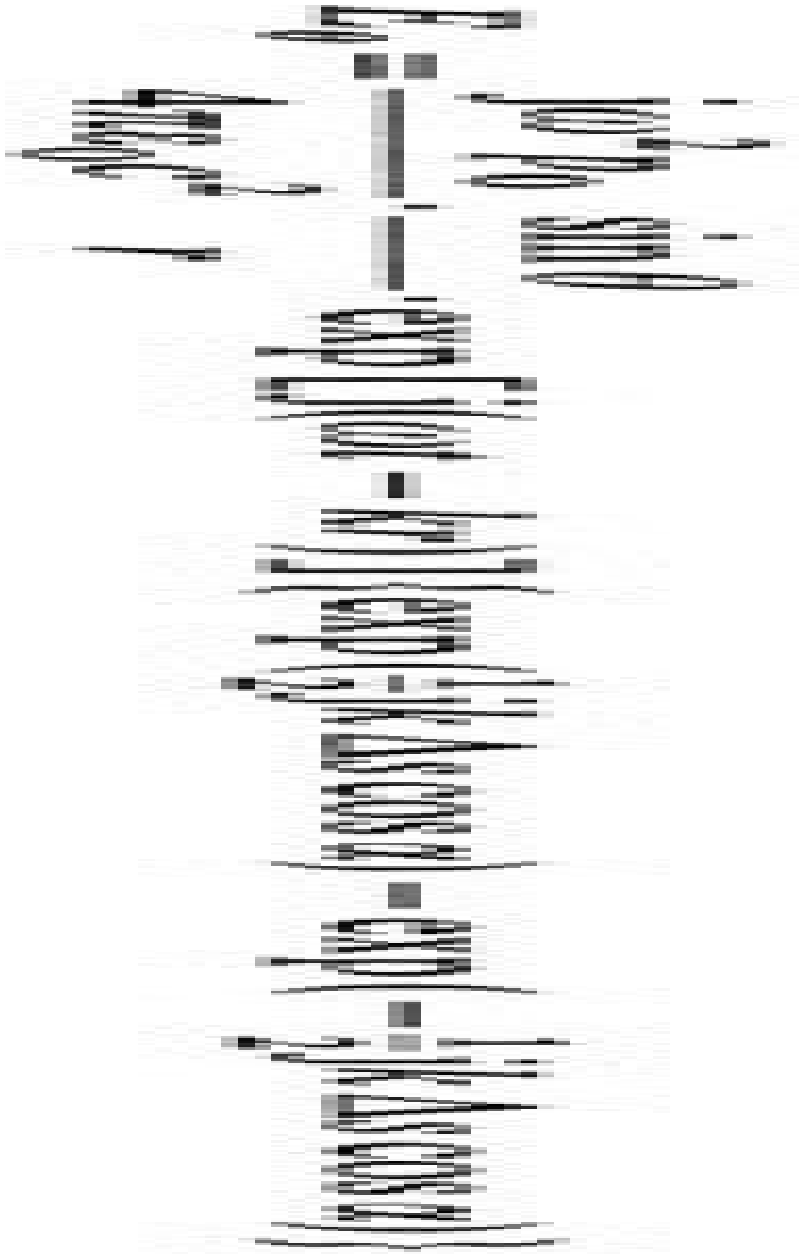
13.4 An antenna of length l shorter than half a wavelength has only a small oscillating current. This is because the current has to be zero at the ends, and the amplitude increases towards the middle according to the formula



One way of making the current larger is to capacitatively load the ends of the aerial, when the current at the ends is no longer zero. This is often done by stringing wires from the top of a vertical antenna. The reflection in the earth is used as the second half of the antenna, and typically a vertical height of ~ 40 m would suffice for a local station with a range of ~ 50 km.

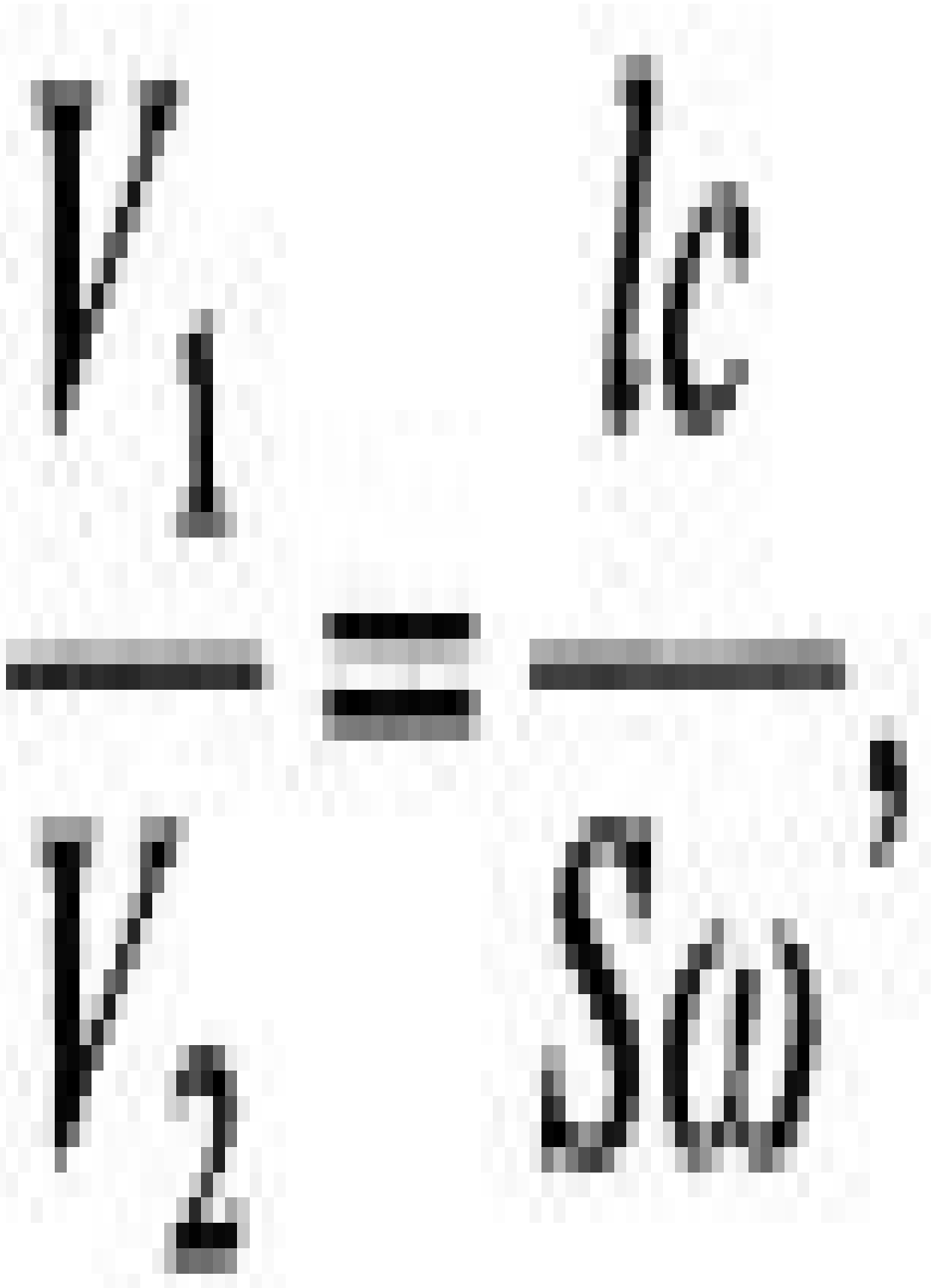
13.5 If the dipoles are separated by a small distance Δs, the electric field at large distances is



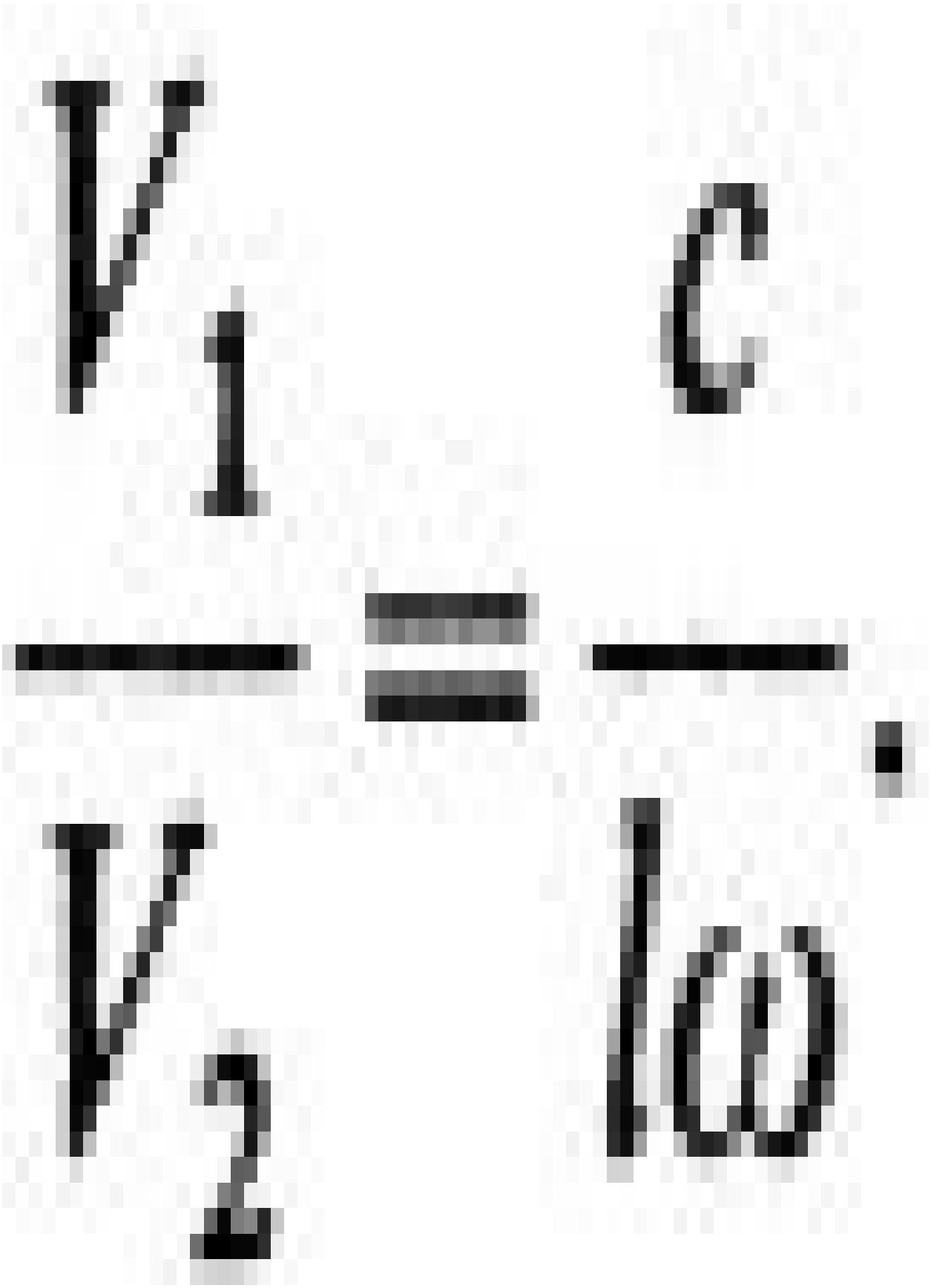


where α is the angle between the direction of observation and the line joining the centre of the dipoles. We have taken the origin to be at the mid point of this line.

13.6 If V1 is the e.m.f. from the short wire and V2 from the loop



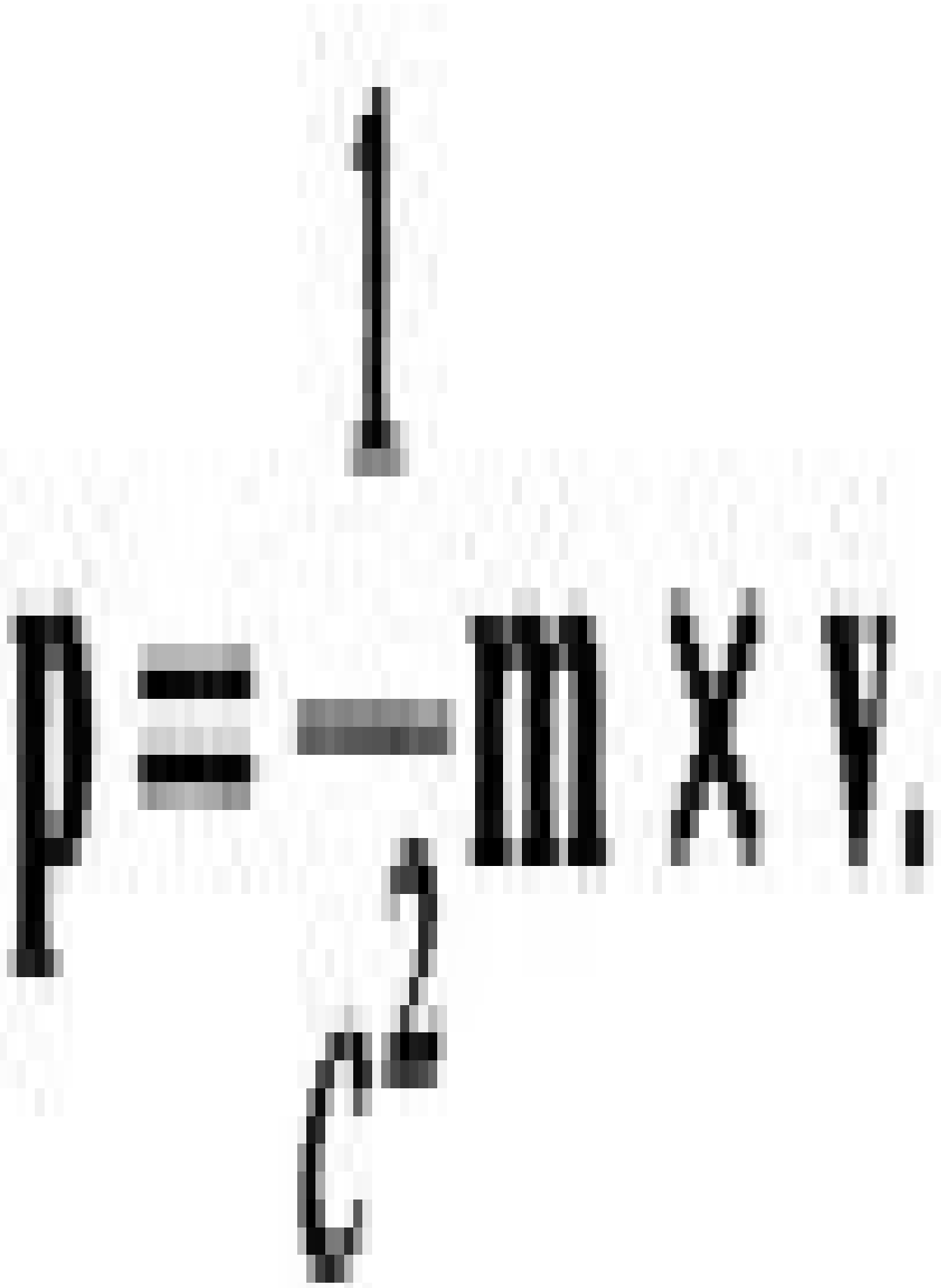
where ω is the angular frequency of the radiation, and c the speed of light. If S ~ l^2,



13.7 The cross-section for combined scattering and absorption is $12\pi(c/\omega)^2$.

PROBLEMS 14

14.1 In the moving frame the wires of length a appear to be charged, and are shortened by the FitzGerald contraction. The magnitude of the coil's magnetic moment is m = lab, and the electric dipole moment in the moving frame is


p = mxv.

14.2 3.2×10^9 V/m.

14.3 The charge at the origin generates no magnetic field along the x-axis. The second charge is acted upon only by the electric field directed along the x-axis. However, the charge at the origin is itself moving perpendicularly to the magnetic field generated by the other charge, and experiences a force in the y-direction. Action and reaction are not equal and opposite, implying that the momentum of the particles is not conserved. This is not as disturbing as it first appears. The fields carry energy and momentum, and they are changing in such a way that momentum is conserved for the whole system of particles and fields.

14.4 The force is readily derived in the same form as in the answer to Problem (4.16). The contribution from two elements dl1 and dl2 to the force is also in a form which does not lead to action and reaction being equal and opposite. However, a steady current can only flow in a complete circuit, and the total forces acting on the two coils are indeed equal and opposite.

Further reading

Below is a short list of books for further reading on topics covered in this book.

B. I. Bleaney and B. Bleaney, Electricity and Magnetism, Vols I and II, Oxford University Press, 1989. This book contains more material on the microscopic aspects of electromagnetism, and on practical applications including electronics.

P. Lorrain and D. R. Corson, Electromagnetic Fields and Waves, W. H. Freeman, 1988, has, as its title implies, a greater emphasis on waves. The companion volume, also by P. Lorrain and D. R. Corson, Electromagnetic Fields and Waves: Principles and Applications, W. H. Freeman, 1990, is laid out in a way to encourage self-teaching and includes many worked examples.

J. D. Jackson, Classical Electrodynamics, Wiley, 1975. This is an advanced text and very comprehensive: a good place to look for discussions ofesoteric and knotty problems.

R. A. Bartkowiak, Electric Circuit Analysis, Wiley, 1985. This book is one of many similar books intended for electrical engineers, and it contains a large number of fairly straightforward worked examples.

F. H. Read, Electromagnetic Radiation, Wiley, 1980. Here Maxwell's equations are used to treat the generation and interactions of radiation, and then quantum mechanical ideas are introduced, allowing a proper discussion of absorption and of lasers.

K. F. Sander, Microwave Components and Systems, Addison Wesley, 1987. Transmission lines, waveguides and antennas are thoroughly treated in this book, which also includes a discussion of sources used to generate microwaves.

C. Kittel, Introduction to Solid State Physics, Wiley, 1986. Now in its sixth edition, this book contains excellent descriptions of the microscopic phenomena underlying the behaviour of materials subjected to electromagnetic fields. Topics included are metallic conduction, dielectric properties, diamagnetism and paramagnetism, ferromagnetism and antiferromagnetism.

G. Stephenson, Mathematical Methods for Science Students, Longman, 1973. Familiarity with the mathematical techniques used in electromagnetism is essential for its proper understanding. Of course there is an enormous number of textbooks at the appropriate level: this one is concise and includes, for example, chapters on vectors, complex numbers and differential equations, all of which are relevant to electromagnetism.

Index

Active circuit elements

Admittance

Alternating current

A.C. networks

complex number representation

Ampere (unit)

Ampère’s law

for the field H

in differential form

in differential form for the field H

Amplitude, of alternating e.m.f.

Angular momentum

of orbiting electron

electron spin

Antennas

Antiferromagnetism

Antinode in standing wave

Armature

Atomic

charge density

currents

electric field

magnetic field

number

Battery

Bending power

Betatron

B–H curve

hard steel

soft iron

BH product

Biot–Savart law

Bohr magneton

Boltzmann constant

Boltzmann factor

Boundary conditions

for B and H

for changing fields

for E and D

Boundary value problems

Branch in network

Brewster angle

Bridge network

Capacitance

in A. C. circuits

of coaxial cable

of parallel cylinders

of parallel plates

stray

Capacitors

Cartesian coordinates

Cavity resonators

Cavity, field in spherical

Characteristic impedance

of coaxial cable

of ladder network

of parallel wires

of strip line

of transmission line

Charge

conservation

density

free

induced

invariance

surface

Circuits see also Networks

equivalent

power in

tuned

Clausius–Mossotti formula

Coaxial cable

capacitance per unit length

characteristic impedance

inductance per unit length

resistance per unit length

Conductance

Conduction

electrons

Conductivity, electrical

table of values

Conductors

plane waves in

Coordinate systems

Coulomb (unit)

Coulomb’s law

Coupling coefficient

Curie temperature

Curl

Current

density

free

grid

loop

magnetization

measurement of

Cut-off frequency

in filter

in parallel plates

in waveguides

Cylindrical polar coordinates

Damping of resonant circuit

Del

Delay line

Depolarizing field

Detector

in slotted waveguide

on transmission line

Diamagnetism

Dielectric

constant

materials

sphere

Differential operators

Differential form of

Ampère’s law

Faraday’s law

Gauss’ law

Dimensions of electrical quantities

Diode

Dipole

Hertzian

potential

see also Electric dipole; Magnetic dipole

Dipole magnet

Dispersion

in waveguides

Displacement current

density

Displacement

electric

Divergence

theorem

Domain

Dynamo

Eddy currents

Electric

charge

dipole

dipole moment

table of values

displacement

quadrupole

susceptibility

table of values

Electric field

atomic

due to dipole

due to ionized atom

flux of

in cavity

induced

in plane wave

lines

local

macroscopic

outside charged conductor

Electrical conductivity

table of values

Electromagnetic waves

in free space

in media

Electromagnetic

energy

induction

radiation

Electromagnets

Electromotive force (e.m.f.)

induced

Electron

gun

in metals

spin

volt (unit)

Electrostatic

energy

energy density

force

images

lenses

potential

shielding

Energy

electrostatic

in resonant circuit

magnetic

of plane wave

released in reactions

Energy density

electrostatic

in changing field

magnetic

Equation of continuity

Equipotential surface

Equivalent circuit

Farad (unit)

Faraday’s law

in differential form

Ferrimagnetism

Ferrite

Ferromagnetic materials

Ferromagnetism

Field

conservative

electric

magnetic

lines

of moving charge

see also Atomic; Local; Macroscopic

Filter

high-pass

low-pass

pass- and stop-bands

band-pass

strip line

Finite differences

Fitzgerald contraction

Flux

electric

magnetic

Fluxmeter

Force

conservative

magnetic

on charged particle

on charged plate

on coils in magnetic fields

on conductors and magnetic materials

on dielectric material

Fourier’s theorem

Four-vectors

Free charge

Free current

Full-wave aerial

Gauss (unit)

Gauss’ law

in dielectrics

in differential form

Gaussian units

Gradient

Group velocity

Half-wave aerial

Hall effect

quantum Hall effect

Helmholtz coil

Henry (unit)

Hertz

unit

Hertzian dipole

Homopolar generator

Hysteresis

losses

Identities

Image charge

Impedance

characteristic

complex

of capacitor

of inductor

in parallel

input

in series

output

Incremental

inductance

permeability

Induced

charges

electric field

e.m.f.

magnetic moment

Inductance

in A. C. circuits

mutual

self

Induction

Inertial frame

Input impedance

of transmission line

of transformer

Insulators

plane waves in

absorption in

Intrinsic magnetic dipole moment

Ionosphere

Inverse square law

precision of

Josephson effect

Joule heating

Kirchhoff’s rules

Klystron

Ladder network

Laplace’s equation

Laplacian operator

in various coordinates

Larmor angular frequency

Leakage flux

Lenses

apertures

magnetic quadrupole

Lenz’s law

Line charges

Line integral

in magnetic field

Lines

of force

of magnetic field

Local field

Logarithmic decrement

Loop

in networks

current analysis

Lorentz

condition

force

invariance

ionization

transformation

Losses

in transformers

in transmission lines

in waveguides

Macroscopic

electric field

magnetic field

Magnetic

circuits

dipole

dipole moment

energy

energy density

flux

force

intensity

mirror

quadrupole doublet

quadrupole lens

rigidity

susceptibility

measurement of

table of values

tape

vector potential

Magnetic field

atomic, local and macroscopic

due to Earth

due to straight wire

field H

in plane wave

inside long solenoid

of small loop

on axis of Helmholtz coil

Magnetic field lines

Magnetization

current

curve

Magnetomotive force

Mass spectrometer

Matching

Maxwell

inductance-capacitance bridge

Maxwell’s equations

Metallic conduction

Microscopic field equations

see Atomic field

Modes

in cavity

in parallel plates

in waveguides

Molecular polarizability

Moment, dipole

electric

magnetic

Monochromatic waves

Motional e.m.f.

Moving charge, force on

Multipole expansion

Mutual inductance

Networks

ladder

parallel

series

Neumann’s formula

Node

in network

analysis

in standing wave

Non-polar liquids

Norton’s theorem

Nuclear

charge

magnetic resonance(NMR)

magneton

Ohm’s law

Ohm (unit)

Optical fibre

Output impedance

of transformer

of transmission line

Owen inductance bridge

Parallel plate capacitor

Paramagnetism

Particle motion in electric and magnetic fields

Passive circuit elements

Period

Permanent magnets

Permeability of free space

Permeability, relative

Permittivity

of free space

relative

table of values

Phase

Phase velocity

see Wave velocity

Phasor

diagram

Photoelectric effect

Photons

Planck’s constant

Plane wave

Plasma

oscillations

Point charge

Poisson’s equation

Polar diagram

Polar molecule

Polarizability

Polarization

by reflection

charge

circular

current

linear

of plane wave

vector

Potential

difference

dipole

electrostatic

four-vector

quadrupole

retarded

vector

Power

factor

in A. C. circuits

radiated

Poynting vector

Primary coil

Propagation vector

Proton

Q (quality factor)

Quadrupole potential

Quantum rules

Quantum theory

of atom

of radiation

Quark

Quarter-wave line

Radiation

field

pressure

resistance

zone

Reactance

Rectifier

Reflection

by conductor

by insulator

coeffcient

on transmission line

Refraction of electromagnetic waves

Refractive index

Relative

permeability

permittivity

table of values

Relativity

Reluctance

Remanence

Resistance

at high frequency

in A. C. circuit

internal

of radiating antenna

Resistivity

Resonance

Retarded potentials

Right hand screw rule

RMS current and voltage

Saturation magnetization

Scalar field

Secondary coil

SI units

Skin depth

Snell’s law

Solid angle

Special relativity

Speed of electromagnetic waves

between parallel plates

in free space

in isotropic media

in waveguides

on transmission lines

definition of

Spherical polar coordinates

Spin of electron

Standing wave

ratio

Stator

Stokes’ theorem

Stray impedance

Strip line

Superconducting

cavities

magnets

Superconductors

Superposition

of electric fields

of magnetic fields

of networks

of potentials

principle of

Surface

charge density

currents

equipotential

integral

Susceptibility

electric

magnetic

T-section filter

TEM waves

TE waves

TM waves

Tesla (unit)

Time dilatation

Thermo-electric e.m.f.’s

Thevenin’s

equivalent circuit

theorem

Tokomak

Torque

on coil

on electric dipole

on magnetic dipole

Transfer function

Transformation

linear

Lorentz

of fields

of forces

Transformer

equivalent circuits

ideal

losses

real

Transients

Transmission lines

coaxial cable

parallel wire

strip line

Transmission coefficient

Uniqueness theorem

Units

Vector

field

potential

Velocity of light see Speed

Volt (unit)

Voltage

standing wave ratio

transfer function

Wave equation

Wavefront

Wavegroup

Waveguide

equation

slotted

Wavelength

in waveguide

Wavenumber

in waveguide

Wave velocity

Weber (unit)

Wheatstone bridge